Author:
Jones, Alexander M

Title:
All You Need are Axioms

A Defence of Deflationism via Formal Truth Theory

All You Need are Axioms: A Defence of Deflationism via Formal Truth Theory

Alexander Marcus Jones

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Arts

School of Arts April 2019

Word Count: 72, 041

Abstract

My thesis presents a novel argument for deflationism about truth using new research in formal truth theory.

In Chapter 2 I show the inadequacy of CT^- as a truth theory for non-standard models of syntax. I develop an extended T-Schema to overcome this and prove it entails non-conservativity over arithmetic. This shows a minimally adequate purely alethic theory has powerful deductive consequences. I argue, against the conservativity argument, this is a boon for deflationism. To do this, Chapter 3 provides a novel account of which theories of truth are deflationary, namely purely logical-linguistic-semantic theories of the word 'true'. Chapter 4 then considers which formal theories of truth are deflationary and argues deflationary theories need not be proof-theoretically conservative, semantically conservative or formally logical. Instead, taking my conception of deflationism from Chapter 4, I argue that all current axiomatic theories of truth are deflationary. In Chapter 5 I develop and explore two new axiomatic theories of truth and argue for one's adequacy as a formal theory of truth. Chapter 6 provides a new philosophical position, deflationary alethic pluralism, which I argue shows even a simple deflationary theory of truth can capture the philosophical and linguistic benefits of a powerful pluralist theory of truth.

I conclude in Chapter 7 that, taken together, these results imply the adequacy of, and provide support for, deflationism about truth. This has important ramifications for those who seek truths. It tells us that theoreticians need not be concerned with metaphysical or epistemic features of truth, particularly those which could conflict with their practice. Further, philosophers using the notion of truth to phrase their theory do not incur additional commitments by doing so. With a deflationary conception of truth, troubling aspects of *truth* are removed from focus, leaving theoreticians free to pursue the *truths* of their particular domain.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

Acknowledgements

Much gratitude is due to my supervisors, Professor Leon Horsten and Professor Daniel Whiting, for their exceptional support and assistance during this process. Leon, your expertise and open-ended questions have guided, sharpened and refined my work immeasurably. Daniel, your breadth and depth of knowledge have offered extremely valuable new insights and perspectives. I could not have asked for better advisers.

I have received great support from staff and students at the University of Bristol, particularly those in the Foundational Studies Bristol research group. Colleagues from beyond Bristol, especially those who have become friends, you have improved the contents of this thesis significantly. Thank you to all who have taken the time to engage with my research and generously provided helpful comments and suggestions. Thank you to Dr. Richard Kaye for inspiring this project, particularly Chapter 2.

This project would not have been completed without generous funding from the South West and Wales Doctoral Training Partnership. I would like to thank them for funding tuition, training and, most importantly for me, the stipend that I've lived on.

Friends and family have been an important source of solace and maintained my motivation. Mum, Dad and Aurea, thank you for always being there and helping me switch my brain off. Grandma, thank you for passing down your love of reading. I think it is fair to say this project would never have been started without it. Siobhan, thank you so much for making these last few months a blessing, rather than a chore. You deserve all the credit for my (comparative!) lack of stress finishing this thesis.

Friends, I am fortunate that there are far too many of you to mention by name, but your contributions have been no less substantive. Fellow Bristol PGRs, your support is incalculable. Cheese and wine has been the best collective therapy I could ask for. Birmingham Brunchers, you have never, ever, failed to entertain. I look forward for the time to read the hundreds of messages now! Finally, and by no means least, everyone who ever offered a supportive word or expressed interest in my work during the past years, thank you.

Table of Contents

Chapter 1

Introduction

The concept of truth appears to be of great importance for understanding our world and is the subject of investigation in this thesis. It seems that one of the goals of personal and societal inquiry is to discover truths. The scientist conducts experiments to try and ascertain truths of the world and the mathematician produces proofs to discover new mathematical truths. The archaeologist uncovers evidence of historical truths and the anthropologist observes social truths. Truth appears to have an important role for living in a developed society also. Legal courts try and determine the truth of a particular accusation and journalists check whether politicians are speaking the truth. Economically, truth seems a guide to profit. The stockbroker analyses records to try and determine the truth of a particular forecast and the marketer conducts research to find what customers truly desire. Truth appears to be at the heart of inquiry and of high importance in understanding and shaping the world.

The concept of truth is ubiquitous in being used to frame these inquiries, but it is highly unclear what this concept is. 'Truth' is used in all the above examples to formulate what each theoretician seeks, but is not itself the object of inquiry. Most inquirers are interested in what the truths are, rather than the concept of truth itself. Yet, this concept is so widespread in its utility for framing inquiry, that it appears to be an important efficacious concept. So what is the concept of truth used in each of the examples above? This is a classical question of philosophical inquiry and the primary question to be explored in this thesis.

I will focus this question by looking at what it means to say, of a truthbearer S, that S is true. Here I use the term 'truthbearer' just to mean something which can

be true or false.[1] I will be assuming that sentences are the bearers of truth, but hopefully other notions of truthbearers, most notably propositions, but also utterances, beliefs, etc. can be substituted in for these without issue. In questioning what it means to say that a sentence S is true I will make a further assumption. I assume that 'is true' is a predicate of sentences which expresses that a sentence S instantiates the truth property. This property is the characteristic or quality that all true sentences have. This thesis will hence be concerned with the behaviour, nature and role of the truth property.

This question has been explored at length throughout the history of analytic (as well as non-analytic and pre-analytic) philosophy. This has resulted in a number of theories of truth that have been formulated and debated. The most prominent of these are the correspondence account, the coherence account, the pragamtist account, and the more recent deflationary and pluralist accounts. The first three of these theories provide monist substantive accounts of truth: they provide theories of a single substantial truth property – they are concerned with *the nature* of truth. A correspondence account of truth states that a sentence is true if and only if it stands in some form of relation to (*corresponds with*) an objective fact in the world.[2] This theory intends to capture the intuition that 'a true sentence expresses the external world as it actually is'. A coherence theory, on the other hand, views a sentence as true if and only if it coheres with an already specified set of sentences.[3] For the coherence theorist, truth does not express a relation between sentences and the world, but sentences and other sentences (often beliefs). A pragmatist theory of truth rejects both these approaches and instead posits truth as a relation between sentences and utility.[4] This has been specified in different ways, but is often expressed by the claim that a sentence is true if and only if it leads to the best success in action or belief. Such theories have contemporary advocates, but each has been criticised on the grounds that its universal nature of truth is in conflict with particular types of true sentences. Whilst these accounts of truth appear promising for some types of truths, the theories do not seem to hold generally.[5]

[1] Soames (1998, Ch. 1) provides a discussion of different notions of truthbearers and what these might be more precisely.

[2] David (2018) provides a recent summary of correspondence theories of truth and their appeal.

[3] Walker (2001), for example, introduces and defends a coherence theory of truth.

[4] Misak (2018) provides an overview of pragmatist theories of truth and their utility.

[5] An overview of such criticisms is provided in Chapter 6.

Pluralism about truth is a contemporary response to this failure and questions one of the key assumptions of these theories – that there is a single truth property.[6] A pluralist theory of truth instead gives a theory of domains of discourse which classifies sentences into different semantic categories. A sentence in one domain of discourse exemplifies a property of truth given by some theory, but this can differ to the truth property exemplified by a sentence in a different domain of discourse. These theories specify that truth has *many* natures. Pluralism about truth will be specified and discussed more fully in Chapter 6.

The final dominant contemporary position is deflationism. Deflationary theories of truth reject the view that there is even a 'substantive' truth property at all. This thesis will be primarily concerned with deflationary theories of truth. These theories claim that truth is *insubstantial* in nature and that the concept of truth is not an explanatory concept. Quite what such claims amounts to is not clear, and this will be discussed in Chapter 3.

Deflationism appears as the most parsimonious of these theories of truth, since it posits only a single insubstantial property of truth. I therefore claim that, everything else being equal, deflationism is the default conception of truth. If deflationism can be shown to be adequate, then it should be held as the correct conception of truth. This thesis will aim to show the adequacy of deflationism and I shall conclude in Chapter 7 that a deflationary understanding of truth is correct.

Traditional debates over the nature of truth have typically been settled, or at least researched, by philosophical argumentation. Parties enter into a dialectic, then valid reasoning and thought experiments are used to refine and critique positions. This thesis will follow this methodology, but will also carry out and use mathematical methods. Since Tarski (1956), the concept of truth has also been investigated using the tools of mathematical logic. This approach uses formal methods to research mathematical objects and principles which aim to describe truth and its properties. This mathematical approach has seen the development of a number of formal theories of truth and a deep understanding of their features.

These theories stand in methodological contrast, if perhaps not philosophical contrast, to the above theories of truth. There are two main categories of formal theories of truth: semantic and axiomatic. They follow in the model-theoretic and proof-theoretic tradition of logic, respectively. Semantic theories of truth aim to

[6]Lynch (2009) details and argues for a plural approach to truth. Other plural theories are available in the literature and these shall be discussed in Chapter 6.

build a class of true sentences, where the true sentences are those satisfied in a particular formal model. Axiomatic theories of truth aim to provide rules for a new truth predicate which has been added to an already established mathematical theory. Both Halbach (2011) and Horsten (2011) provide comprehensive and detailed overviews of these and Cieśliński (2017) provides an up-to-date description of some of the most recent results in the field. A brief technical introduction to these formal theories of truth can be found in Section 1.2. The philosophical impact of this recent logical research, and the connection between formal theories of truth and the philosophical theories of truth mentioned above, is still under much debate. This thesis provides a novel development of this research.

One main body of work connecting formal theories of truth and philosophical theories of truth is in the exploration and proposed resolution of semantic paradoxes. Field (2008) provides an influential discussion of much recent work in this area and the debates that are ongoing. This thesis will, for the most part, avoid discussing the paradoxes, with Chapter 5, Section 5.2.5 being the exception. The other large body of work connecting formal theories of truth with philosophical theories of truth is research about deflationism and this thesis sits comfortably within this research programme.

Formal theories of truth have been used to criticise, support and formulate deflationary theories of truth. One main argument in this area is the 'conservativity argument' against deflationism. According to this argument, proposed by Horsten (1995), Shapiro (1998) and Ketland (1999), formal theories of truth allow us to prove new mathematical theorems, and hence truth can be explanatory, in contradiction with deflationary claims. The argument has been widely debated by authors such as Field (1999), Tennant (2002), Cieśliński (2007, 2010a,b, 2015), Nicolai (2015), Galinon (2015), Horsten and Leigh (2017) and Fujimoto (2019). A development of this argument will be the subject of Chapter 2 and the conservativity argument shall be further discussed in Chapter 4. The correct formulation of deflationism has been investigated by authors such as McGee (1992), Halbach (1999, 2011) and Picollo and Schindler (2018) and I will discuss my take on this in Chapter 4 as well. Arguments for deflationism using formal truth theories are harder to find and are most notably provided by Horsten (2011) and Cieśliński (2017), who provide arguments for particular deflationary theories of truth using their favoured formal theories of truth. I will conclude my thesis (Chapter 7) with a new argument for deflationism using formal theories of truth.

My thesis will develop this existing research connecting formal theories of truth and deflationary theories of truth. One primary research question is whether formal theories of truth support or oppose deflationary theories of truth. As the brief summary of research above suggests, much current work in this area is provided by deflationists who try and defend their view from the conservativity argument. In fact, a brief summary of the literature would suggest that deflationism has been strongly challenged by the results of those working in formal theories of truth. One of the main contributions of my thesis is to push against this suggestion, and instead argue that we should see work in axiomatic theories of truth as explorations of deflationism about truth.

I will provide a novel conception of deflationism (Chapter 3) which leads to a new argument that our current axiomatic theories of truth are deflationary theories of truth (Chapter 4). In order to do this I will provide a novel strengthening of the conservativity argument, which blocks prominent responses (Chapter 2), but will go on to argue that conservativity is not a commitment of deflationism (Chapter 4). I suggest that instead the important question to focus on is whether any axiomatic theory of truth adequately captures the concept of truth. To begin to provide an answer to this question I develop and explore two new axiomatic theories of truth (Chapter 5) and argue that one of these has a number of attractive features as a formal theory of truth. I then provide a new philosophical position, deflationary alethic pluralism (Chapter 6), which I argue shows that even a simple deflationary theory of truth can capture the philosophical and linguistic benefits of a powerful pluralist theory of truth. Together, I take these arguments to provide a new argument that formal theories of truth support a deflationary conception of truth. I will thus conclude (Chapter 7) that deflationism is the most appropriate conception of truth.

This result has important ramifications for those seeking truths. In the examples provided above, I phrased the various theoreticians' inquiries using the concept of truth. A deflationary notion of truth tells us that this is all truth is for – assisting with this phrasing, and that there is no deep substantive nature to truth beyond this. Those seeking truths are not searching for anything special to do with the property of truth, other than a fragment of its extension. This tells us that theoreticians do not need to be concerned with particular metaphysical or epistemic features of truth, particularly those which may offer conceptual or practical conflict with their practice. Further, this means that philosophers using the

notion of truth to phrase a particular theory do not incur any additional commitments or baggage by doing so. For example, ethicists are at liberty to endorse that there are moral truths, without endorsing that there are worldly moral facts – as a correspondence theory of truth would commit us to. Metaphysicians can endorse that some ultimately unknowable sentences still have a truth value – against most coherence theories of truth. With a deflationary conception of truth, troubling aspects of *truth* are removed from focus, leaving theoreticians free to pursue the *truths* of their particular domain. Such benefits of deflationism are discussed in more detail in Chapter 7.

In the following section I shall provide a more detailed overview of the chapters of this thesis, and then in Section 1.2 I shall detail some of the common technical notions used throughout this thesis.

1.1 Thesis Contents

The chapters of this thesis are intended to stand alone as much as possible. Each has been written without requiring knowledge of the other chapters, or the overall argument running through the thesis. Each is also of independent interest and can be appreciated without regard to the wider thesis. Further, each chapter contains its own summary of the literature relevant to that chapter. This does mean that there will be occasional repetition of concepts across chapters, although formal notions and abbreviations common to all can be found in the following section, Section 1.2. Whilst this means the chapters do not need to be read in order, it would certainly benefit the reader to do so. Each chapter begins with a brief mention of how it fits into the underlying structure of the thesis and develops on arguments from the previous chapters. In this section I will provide an overview of each chapter and discuss the overlying argument running through the thesis.

The thesis begins by exploring the T-Schema – a schema of the form 'σ' is true if and only if σ, where 'σ' ranges over truthbearers. Conformity with this schema is commonly understood as a minimal adequacy condition for both philosophical and formal theories of truth. For formal theories of truth, this is interpreted over the standard model of arithmetic, $\mathbb{N}$. I argue in Chapter 2, however, partially due to the possibility of deflationism, that we should consider this over nonstandard models of arithmetic[7] also. I show how we can define an extended T-Schema

[7]These are consistent mathematical models of our agreed axioms of arithmetic, but which

which interprets the T-Schema for nonstandard models of arithmetic as well. I prove that when this is done a simple conservative theory of truth (CT^-) is no longer minimally adequate. I further prove that if we close this theory under the extended T-Schema, then we have a non-conservative theory of truth. I discuss the ramifications that this has for deflationism in light of the conservativity argument and propose that the deflationist has a dilemma: they can accept conservativity and argue against considering nonstandard models, or reject conservativity and argue that the T-Schema has more deductive power than may initially be thought. This dilemma is resolved in Chapter 4 – I propose that to understand whether deflationists should be threatened by, or embrace, nonconservativity requires a better understanding of what deflationism means philosophically and formally.

In chapter 3 I take up this question from the end of Chapter 2 and explore what it means philosophically to be a deflationary theory of truth. I first consider the common understanding that a deflationary theory of truth is a theory which takes some form of T-Schema to be all there is to truth and provide a number of examples and arguments against this equivalence. The chapter then considers current proposals of what it means for a deflationary property of truth to be 'insubstantial'. I provide counterexamples against these proposals and argue these are inadequate also. I propose a new understanding of deflationism, according to which a deflationary theory of truth is a logical-linguistic-semantic theory of the word 'true'. Here I understand this to mean a theory for which a description of the word 'true' using solely logical, linguistic and semantic notions *exhausts* our understanding of truth. I argue that this adequately categorises our current theories of truth and is hence a good understanding of deflationism. I then return to the question of what makes a deflationary truth property insubstantial and argue that it is because these properties are *pleonastic* properties in the sense of Schiffer (2003). This provides a good conception of what deflationism is which can be used to assess the status of arguments like the conservativity argument and to assist in answering the question of which formal theories of truth are deflationary.

Chapter 4 is focussed around this question of which formal theories of truth, if any, should be regarded as deflationary. I first explore the status of the conservativity argument and argue that deflationists should not commit to conservativity, since this establishes a false equivalence between deductive power and explanat-

contain 'nonstandard' numbers larger than the standard natural numbers. These are specified more precisely in Chapter 2.

ory power. I then consider a variant conservativity argument, which argues that deflationary theories of truth are semantically conservative. A theory T is semantically conservative over B if all models of B can be expanded to models of T. This is argued to be an explication of metaphysical insubstantiality, but I argue against this interpretation. I then investigate whether formal truth properties can be regarded as logical properties (and hence deflationary) and conclude that whilst informative such analyses cannot capture all formal theories of truth which are deflationary. This is because most deflationists do not regard truth as solely logical. I argue that, using my criterion of Chapter 3, instead we should regard all current axiomatic theories of truth as deflationary theories of truth, since all of these are logical-linguistic-semantic in nature. This tells us that, to assess whether deflationism is correct using formal truth theory, we should be looking at whether any of our axiomatic theories of truth are suitably adequate as theories of truth – whether they can explain all the features of truth they need to. Here I distinguish two notions of adequacy: formal adequacy, that the theory captures the expected behaviour of the truth predicate, and philosophical adequacy, that it can capture common philosophical uses of truth.

In chapter 5 I develop and explore two new axiomatic theories of truth as part of my assessment of whether any axiomatic theories of truth are adequate. The first of these, Axiomatic Typed Truth (ATT), is aimed at developing a typed theory of truth (a theory of truth where the truth predicate is not self-referential) which overcomes weaknesses of previous typed theories. I show that this theory has a number of attractive formal features and demonstrate how it answers semantic paradoxes. I thus posit it as a formally adequate axiomatic theory of truth. I also exhibit the theory's relation to current debates on quantification and absolute generality and highlight this as a new area of interesting research for formal truth theory. The second part of the chapter uses the internal logic of ATT to motivate and develop a new axiomatic theory of truth which are type-free (the truth predicate can be self-referential). I prove a number of formal properties about these theories and explore the axiomatic theory's connection with a prominent type-free theory of truth (KF) and with ATT. I conclude with some remarks on what this shows about the relation between typed and type-free theories of truth and argue that ATT is adequate as a formal theory of truth. Whilst this shows that we have an axiomatic (deflationary) theory of truth adequate for formal reasoning, the question remains whether an axiomatic treatment of truth is adequate for our

philosophical and linguistic purposes.

Chapter 6 explores this question by arguing for the adequacy of a simple deflationary theory of truth consisting of a type-free T-Schema. If even a very weak deflationary theory of truth can be shown to be adequate philosophically, then a strong theory of truth such as ATT shall be as well. I analyse recent pluralist theories of truth which admit a number of (potentially) substantive truth properties and use these to explain key properties of certain domains of discourse. A frequent objection to deflationism is that it is not able to account for this explanatory aspect of truth. I argue against this, however, and propose that even a simple deflationary theory of truth can admit truth-like properties which can perform the same role as plural truth properties. I call this position deflationary alethic pluralism and show that it both defends deflationism from challenges of inadequacy and is not beset by typical challenges facing pluralism about truth. I argue that this allows the deflationist to claim a theory of truth which accommodates common philosophical uses of truth and thus a philosophically adequate theory of truth.

I conclude (Chapter 7) with my overall argument for a deflationary conception of truth. I argue that this should be the *default* conception of truth, since it is ontologically the lightest, and thus it needs only to be shown to be adequate. I argue that from Chapter 3 we should understand deflationism to be the view that all we need for a theory of truth is a logical-linguistic-semantic theory of the word 'true' and that therefore (from Chapter 4) all axiomatic theories of truth are deflationary. I then argue that we have an axiomatic theory of truth, ATT, which is adequate both formally (from Chapter 5) and philosophically (from Chapter 6). I conclude that this leads to the conception of deflationism as the correct conception of truth. I end by discussing the importance of this result for philosophers and theorists from other disciplines and note further research questions inspired by this thesis.

1.2 Common Technical Notions

Before moving onto the first chapter of this thesis, I shall introduce many of the formal notions and notation that I will use throughout this thesis. I shall assume familiarity with common techniques and concepts from mathematical and philosophical logic, all of which should be contained within a standard textbook. Boolos, Burgess and Jeffrey's (2007) *Computability and Logic*, for example, is more

than sufficient. I shall be working within a classical suitably strong metatheory, such as ZFC, and will assume familiarity with common set-theoretic notation and notions, including that of ordinal and cardinal numbers. The formal focus of my thesis shall predominantly be concerned with expanding the theory of first order Peano Arithmetic with truth predicates and exploring how we can define these.

Peano Arithmetic (PA) is a theory in the language of arithmetic $\mathscr{L}_A$. Each language I consider will contain the standard first order connectives $\neg$, $\wedge$ and $\vee$,[8] quantifiers $\exists$ and $\forall$, the identity relation $=$, an infinite set of variables $x_0, x_1, \ldots$ and brackets (and) as punctuation symbols. In addition to these, the language of arithmetic contains the constants 0 and 1 and binary relations $<$, $+$ and $\cdot$. I shall write $\bigwedge_{i<n} \varphi_i$ and $\bigvee_{i<n} \varphi_i$ as shorthand for the repeated conjunction and disjunction of a set of formulas φ_i for $i < n$ respectively. More explicitly as an example $\bigwedge_{i<n} \varphi_i$ is shorthand for the formula:

$$(\varphi_1 \wedge (\varphi_2 \wedge (\varphi_3 \wedge (\ldots \wedge (\varphi_{n-2} \wedge \varphi_{n-1}) \ldots))))$$

I will frequently classify formulas of arithmetic within the arithmetical hierarchy. This is defined inductively over natural numbers. The class of $\Delta_0 = \Sigma_0 = \Pi_0$ formulas are formulas from $\mathscr{L}_A$ which contain only bounded quantifiers. The class of Σ_{n+1} formulas is the class of all formulas equivalent to $\exists x \varphi$ where φ is a Π_n formula. The class of Π_{n+1} formulas is the class of all formulas equivalent to $\forall x \varphi$ where φ is a Σ_n formula.

Peano Arithmetic is a powerful theory of arithmetic which consists of the axioms of a discretely ordered semiring (PA^-) together with an induction scheme for all formulas φ in $\mathscr{L}_A$. The standard model of PA is the natural numbers $\mathbb{N}$, but it should be noted that the theory admits nonstandard models as well. For more details on this theory, its models and its features the reader is referred to, for example, Kaye's (1991) *Models of Arithmetic*. A summary of the most salient features of this theory for the thesis are provided below.

Peano Arithmetic is able to encode formulas of $\mathscr{L}_A$ as numbers and perform many syntactic and mathematical operations on these. If $\varphi(\bar{x})$ is a formula in the language of arithmetic, then its Gödel code will be denoted by $\ulcorner \varphi(\bar{x}) \urcorner$, where $\bar{x}$ represents a finite tuple of free variables. We can define a substitution function $Subs(\ulcorner \varphi(\bar{s}) \urcorner, \bar{t})$ which returns $\ulcorner \varphi(\bar{t}) \urcorner$ (the substitution of $\bar{t}$ for $\bar{s}$) where φ is an

[8] I will often refer to the additional standard 'connectives' $X \to Y$ and $X \leftrightarrow Y$, but technically these are abbreviations for the formulas $\neg X \vee Y$ and $(X \to Y) \wedge (Y \to X)$ respectively.

$\mathscr{L}_A$-formula and $\bar{s}$ and $\bar{t}$ are tuples of terms. PA can define many useful syntactic functions such as: *Term*(x), *ClTerm*(x), *Sent(x)*, *Form(x)* and *At*(x) which express that x is the Gödel code of a term, closed term, sentence, formula and atomic sentence of $\mathscr{L}_A$ respectively. Other useful functions that PA can define[9] are *Val*(x), which calculates the value of the (closed) term x codes, and '$\dot{\cdot}$' functions for each connective and the negation symbol, which return the code of the connective or negation applied to its argument(s). For example, given formulas x and y, the function $\dot{\wedge}(\ulcorner x \urcorner, \ulcorner y \urcorner)$ returns $\ulcorner x \wedge y \urcorner$ and $\dot{\neg}(\ulcorner x \urcorner)$ returns $\ulcorner \neg x \urcorner$. There are also '$\dot{\cdot}$' functions for the quantifiers as well, for instance, for a formula $\varphi(\bar{x})$ and tuple of variables $\bar{t}$ we have that $\dot{\forall}(\ulcorner \varphi(\bar{x}) \urcorner, \bar{t})$ returns $\ulcorner \varphi(\bar{t}) \urcorner$. I also include '$\dot{\cdot}$' functions for the relations $=$ and $<$. For example, given closed terms x and y, the function $\dot{=}(x, y)$ returns the code of the sentence that the objects denoted by x and y are equal. If these functions are not applied to codes of formulas, then they return 0, and I will also take $\dot{\rightarrow}$ to be well-defined in the natural way, even though it is not strictly a connective in our setting.

This '$\dot{\cdot}$' notation will be extended to mapping numbers n to their numerals $\hat{n}$, denoted with $\dot{\cdot}$ over the argument. This will be used to bind variables. For example, given a property P, an expression such as $\forall x P(\ulcorner \varphi(\dot{x}) \urcorner)$ is shorthand for $\mathit{Subs}(\forall x P(\ulcorner \varphi(x) \urcorner), \hat{x})$. This means that (the code of) φ satisfies P for every numeral $\hat{x}$. Halbach (2011, p. 32) provides a more detailed explanation of how such notation is used and PA's ability to provide these syntactic utilities. Such coding details quickly become unwieldy, and following convention I will frequently be informal about such details. I will often use a convenient shorthand or refer to a formula rather than its code. Context should make it clear what is meant in such cases and this will significantly increase readability.

That Peano Arithmetic has access to such syntactic operations of its own language results in powerful metamathematical expressibility. For example, PA can define a provability predicate $\mathit{Prov}_{\mathrm{PA}}(x)$ which satisfies Löb's Derivability conditions (Löb, 1955) to express that the theory PA can prove x. This means that PA can also express its own consistency ($\mathit{Con}(\mathrm{PA})$) as the formula $\neg \mathit{Prov}_{\mathrm{PA}}(0 = 1)$, although notably it cannot prove this formula as shown by Gödel's Second Incompleteness Theorem (Smith, 2007).

This thesis will be interested in the addition of truth predicates $\mathit{Tr}(x)$ to Peano

[9]Peano Arithmetic can represent all recursive functions and these are primitive recursive. For details and proofs of this, see Kaye (1991), for example.

Arithmetic. PA is able to define 'partial' truth predicates Tr_S which define truth for strict subclasses S of the language of arithmetic. What is meant by truth predicate here is that it satisfies Tarski's (1956) material adequacy condition for all sentences σ in S:

$$\mathrm{PA} \vdash \sigma \leftrightarrow Tr_S(\ulcorner \sigma \urcorner)$$

Most notably PA defines partial truth predicates Tr_{At} for atomic formulas and Tr_X for classes X of the arithmetical hierarchy. Tarski's famous theorem on the undefinability of truth (Tarski, 1956) states that PA cannot define such a truth predicate when S is the whole language of arithmetic $\mathscr{L}_A$.[10]

This thesis will be concerned with truth for the whole language of arithmetic and will hence expand $\mathscr{L}_A$ to a new language $\mathscr{L}_{Tr} = \mathscr{L}_A \cup \{Tr\}$ where Tr shall be a new predicate symbol for truth. When required, the syntactic operations that PA provides above, such as expressing that a given number x codes a formula of $\mathscr{L}_{Tr}$, will be naturally extended to the language $\mathscr{L}_{Tr}$. We will explore theories of truth T which provide axioms governing this new truth predicate symbol. These are known as axiomatic theories of truth. Many axiomatic theories of truth T have already been developed and this thesis will assume familiarity with a number of these. As such, precise details of each theory shall not be provided here, but will be offered within individual chapters when focussed upon. Details and properties of all the theories I shall mention can be found in a standard textbook on the subject, such as Halbach's (2011) *Axiomatic Theories of Truth*.

One main distinction within axiomatic theories of truth is the typed theories and the type-free theories. Informally, a typed theory of truth disallows the application of the truth predicate to sentences containing that same truth predicate. A type-free theory of truth, on the other hand, generally has no restriction on the self-applicability of the truth predicate.[11] Standard typed theories of truth are the theories of TB – Tarski Biconditionals, CT – Compositional Truth and RT – Ramified Truth. Standard type-free theories of truth are the theories FS – Friedman-Sheard (Friedman and Sheard, 1987), KF – Kripke-Feferman (Feferman, 1991) and PKF – Partial Kripke-Feferman (Halbach and Horsten, 2006). Such truth theories T contain an induction scheme for all formulas in $\mathscr{L}_{Tr}$, but

[10] Tarski's Theorem is of course much stronger than this and states that no sufficiently strong theory can define a truth predicate for the whole of its language.

[11] Precisely setting out this distinction is not so straight forward as it might appear - see Halbach (2011, §10) for example.

we can also consider variant theories T^-, which represents the theory of truth T without induction axioms for $\mathscr{L}_{Tr} \backslash \mathscr{L}_A$.

We will also consider a different kind of formal theory of truth, the semantic theories of truth. These are model-theoretic interpretations of our truth predicate Tr – constructions of the class of true sentences. Again we have a distinction between typed semantic theories of truth, where members of the class cannot code sentences using the predicate Tr, and type-free theories, where members may. Tarski's theory of truth (1956) is the most well-known typed semantic theory of truth, and this can be used to build Tarski's hierarchy of truth predicates Tr_{n+1} which define the class of true sentences from $\mathscr{L}_{Tr_n}$. Kripke's theory of truth (1975) is the most well-known type-free semantic theory of truth, but there is also Gupta and Belnap's revision theory of truth (1993). For more details on these theories and their constructions Halbach's (2011) *Axiomatic Theories of Truth* is again a comprehensive resource.

These technical notions should be sufficient to start the contents of this thesis. I begin the thesis with an exploration of the typed axiomatic theory of truth CT^- and its suitability over nonstandard models of arithmetic. The chapter shall contain its own introduction to the theory CT^- as well as information on the nonstandard models of arithmetic. This will prove to have interesting consequences for the adequacy of deflationism, the conservativity argument and motivate much of this thesis's subsequent research.

Chapter 2

Nonstandard Syntax vs. the T-Schema

One of my primary research questions is the correctness of a deflationary conception of truth and whether formal truth theory can help to answer this. This chapter shall research this question by formally examining the T-Schema and discussing its consequences for deflationism. Deflationists commonly endorse a T-Schema as an essential feature of truth and, as will be argued against in Chapter 3, the T-Schema is sometimes even taken to be definitional of deflationism about truth. This chapter will show formally that a T-Schema can provide more deductive power than initially thought. It is unclear whether this research in formal theories of truth supports or opposes deflationism about truth and this will motivate both Chapter 3 and Chapter 4 where it will be decided that this offers support to deflationism about truth.

Chapter Abstract

In this chapter I propose and discuss a T-schema for nonstandard models of syntax. I show that weak theories of truth which satisfy this are nonconservative over Peano Arithmetic and discuss the ramifications this has for deflationism. The typed T-schema is widely agreed as necessary for any theory of truth, but is studied formally with an implicit assumption of a standard model of syntax. I shall question this assumption and argue that it is not well-motivated, particularly from a deflationary point of view. I explore the inadequacy of compositional truth without induction axioms (CT^-) for nonstandard models of arithmetic and show that we can overcome this by introducing an 'extended T-schema' for nonstandard

models. I prove that closing the theory under this extended T-schema can prove the consistency of PA. This shows that nonconservativity of adequate truth theories arises solely due to alethic considerations, strengthening the conservativity argument against deflationism. On the other hand, if the deflationist accepts nonconservativity, then this shows the deductive power that a T-schema can provide.

2.1 Introduction

Within the study of theories of truth, both formally and informally, the T-schema is a fundamental principle that a theory of truth must satisfy. This is the schema:

$$Tr(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

for all sentences φ of the language under consideration and an appropriate truth predicate Tr. It is also known as the Disquotational Schema, Equivalence Schema and the Tarski-Biconditionals, but I shall refer to it as the T-Schema (TS).[1]

The typed version of this principle can be thought of as a necessary condition for a minimally adequate theory of truth. Take a language $\mathscr{L}$ and some theory T in the language $\mathscr{L}$. A theory of truth for T, in the language $\mathscr{L} \cup \{Tr\}$, should entail all instances of the schema for the language $\mathscr{L}$. This is a typed theory of truth in the sense that, as per Tarski, the truth predicate cannot refer to itself. This means that the theory of truth blocks the Liar Paradox and other problematic sentences. The T-schema for the truth-free language is uncontroversial and should be taken as a minimally adequate condition for a theory of truth to satisfy.

Support for a T-schema is widespread among philosophers of truth. Armour-Garb and Beall (2005, p. 2) write:

> "At least since Tarski (if not since Aristotle) most philosophers have taken the following T-schema to be central to our concept of truth ... Both deflationists and substantivists acknowledge the centrality of this schema."

Substantivists here are those who endorse a traditional theory of truth which aims to describe the *nature* of truth, in opposition to deflationists. Many deflationists

[1] This difference in terminology is due to different notions of truthbearers. I will discuss this and the T-schema in more detail in Chapter 3, Section 3.2.

even go so far as to endorse some kind of (untyped) T-schema as their sole theory of truth. Horwich (1998) endorses a variant of this, for example, but treats propositions as the bearers of truth.

Logicians working on formal theories of truth similarly hold the T-schema as a minimal condition that a formal theory of truth satisfies. Halbach (2011, p. 19), for example, writes:

> "That a definition of truth has the equivalences [T-schema] as consequences is a necessary condition for its adequacy in the intuitive sense"

Whether one investigates theories of truth formally or philosophically, the T-schema is inescapable as a foundational guide and necessary implication. In their application of the T-schema, however, both groups implicitly assume a standard model of syntax, in a formal sense. Within this chapter I shall question the assumption of a standard model of syntax and argue that, particularly in the formal case, this assumption is not well-motivated and should be dropped. I will then explore the unintuitive behaviour that formal theories of truth, in particular the Tarskian compositional axioms CT^-, can produce when nonstandard syntaxes are considered. I show that this can be grappled by a modified version of Robinson's (1963) definition of semantics for nonstandard languages. This can be thought of as providing an extended T-schema that is appropriate for nonstandard models of syntax also.

I discuss the ramifications of this extended T-schema for the 'conservativity argument' against deflationism, which argues that a deflationary theory of truth must be proof-theoretically conservative. I show that closing CT^- under this new extended T-schema proves induction for Δ_0 formulas in the language with the truth predicate and thus a nonconservative theory of truth. This leads to the philosophical conclusion that deflationists who desire conservativity must include, and argue for, the assumption of a standard model of syntax with their theory of truth, for otherwise I show their theory can not be both conservative and entail this extended T-schema. On the other hand, for deflationists who do not require conservativity, this provides formal support and argumentation for the deductive work that a T-schema can provide. This choice is motivational for much of this thesis and shall be resolved in Chapter 4.

2.2 Nonstandard Syntax

In the subsequent sections I shall examine the T-schema formally, in order to clarify the discussion and bring formal methods to bear. I will then reconvene at the end to reflect upon how this impacts philosophy of truth. In order to examine it formally, I shall take the first order theory of Peano Arithmetic as my base theory. This is a suitable background theory since it provides a well-known and powerful theory of arithmetic which is both sound, complete and recursively axiomatised. Further, it is sufficiently powerful to encode statements in the language of arithmetic as numbers within an interpretation by utilising Gödel coding.

2.2.1 Technical Setup

I shall take as my original language the language of arithmetic $\mathscr{L}_A$ and I shall take as my background theory the theory of Peano Arithmetic (PA). Models of this theory shall be denoted by capital letters $M, N, \ldots$ and henceforth, unless qualified further, a model shall refer to a model of PA. I shall often denote the domain of a model as the model itself, and write $a \in M$ to mean a is an element of the domain of a model M. I will denote by $\mathscr{L}_A(M)$ the language $\mathscr{L}_A \cup \{a : a \in M\}$. The natural numbers $\mathbb{N}$ are a model of Peano Arithmetic, but the theory generates other models as well. These are known as nonstandard models and are end-extensions of the standard model $\mathbb{N}$. If M is a nonstandard model and $a \in M \backslash \mathbb{N}$ then a is known as a nonstandard number, since a is larger than any standard natural number, but still finite in the sense of M.

We know that there are nonstandard models of PA of all infinite cardinalities thanks to the Löwenheim-Skolem theorem, but I will restrict my interest only to the countable models. The reason for this is largely practical, in that the countable nonstandard models of PA have the nice structure of $\mathbb{N} \cup (\mathbb{Z} \times \mathbb{Q})$ and that much of the model theory of the particular theory of truth I shall explore is developed over countable models. Henceforth, when I refer to a nonstandard model of arithmetic I thus mean a countable nonstandard model of arithmetic.

Peano Arithmetic is powerful enough to express sentences in the language of arithmetic as numbers via Gödel coding and to encode syntactic operations and functions. Of relevance to this chapter will be the atypical functions *SetSent*(x), *QuantSeq(x)*, and *SkForm*(x, y). The former of these expresses that x codes a set of Gödel codes of sentences and *QuantSeq(x)* expresses that x codes a well-formed

sequence of quantifiers and their variables. $SkForm(x, y)$ expresses that y codes a formula in Skolem Normal Form which is equivalent to the formula that x codes. We note that this is possible since PA has definable Skolem 'terms' – if $\varphi(\bar{x}, y)$ is an $\mathscr{L}_A$-formula, then PA defines the least y such that $\varphi(\bar{x}, y)$ holds.[2]

One interesting facet of nonstandard models is that they carry their own interpretation of PA's theory of syntax. In particular, as a corollary of the 'overspill principle'[3] there are nonstandard numbers a, b, c in any model $M \vDash \text{PA}$ where $M \ncong \mathbb{N}$ such that $M \vDash Term(a)$, $M \vDash Sent(b)$ and $M \vDash Form(c)$. Any sentence with a finite (in the sense of $\mathbb{N}$) length, containing only standard terms and variables, will have a Gödel code within $\mathbb{N}$. Therefore, a nonstandard model of PA will interpret these numbers a, b and c, where b and c only reference standard terms and variables, as being the Gödel codes of terms, sentences and formulas with length not equal to any natural number. A nonstandard model of Peano Arithmetic carries with it an interpretation of its syntax wherein sentences and formulae, in particular their number of connectives, quantifiers and terms in a relation, can have the length of any element within its domain, including nonstandard numbers. I will denote the language of nonstandard sentences including a constant symbol for each element of the model's domain by ${}^*\mathscr{L}_A(M)$.

Some of these nonstandard sentences can be visualised easily. Recursive definitions of sentences, such as $\varphi_0 = (0 = 0)$ and $\varphi_{n+1} = (\varphi_n \wedge (0 = 0))$, provide examples of nonstandard sentences, when n is allowed to range over all elements of a nonstandard model's domain. Sentences with nonstandard indices are infinitely long in length, from our outside perspective, and not well-founded in structure, but are treated as ordinary sentences of arithmetic by the model.

I shall expand the language of arithmetic $\mathscr{L}_A$ to an extended language $\mathscr{L}_{Tr} = \mathscr{L}_A \cup \{Tr\}$ where Tr is intended to be a truth predicate. Due to reasons of Tarskian undefinability, this is taken to be a strict expansion of the language of arithmetic. I will explore various interpretations of the truth predicate and the details of this are fleshed-out in Section 2.3.1.

[2]It should be noted that these are not strictly terms in our language, but for ease of readability we will treat them as if they are. The author is only aware of the details of this procedure provided in an unpublished work by Kaye (2012).

[3]This principle implies that, for nonstandard models of arithmetic M, if $M \vDash \varphi(a)$ for all $a \in \mathbb{N}$, then $M \vDash \varphi(b)$ for some $b \in M$ where b is nonstandard. More details on this principle can be found in a standard textbook such as Kaye's (1991, §6.1).

2.2.2 Considering Nonstandard Syntax

In subsequent sections I shall be examining the behaviour of a theory of truth over nonstandard sentences, but first I will argue that this is a worthwhile avenue to explore. Their study is motivated for three main reasons: a desire to minimise assumptions, consideration of the sceptical position that we cannot know we are not in a nonstandard model of syntax and the utility that nonstandard models can provide to our knowledge of standard models.

When exploring minimal adequacy conditions on a theory of truth, it seems highly desirable to minimise the number of commitments that must be made from a metatheoretic perspective. I do not want to make strong assumptions about how a theory of truth should be, before setting out on an exploration of truth. The exploration of truth might well provide many metatheoretic commitments about how a theory must be, but I do not want to build these in at the start from ad hoc reasons. This is motivated from giving acknowledgement to deflationary positions about truth, where truth is insubstantial and does not play a 'causal-explanatory role'; often truth is referred to as a purely logical notion.[4] Whilst it is not entirely clear what these claims should refer to in a formal context,[5] the fewer metatheoretic commitments made about the requirements of a theory of truth is so much the better for the deflationist. Indeed, Horsten (2011, p. 20) writes "Most of them [contemporary deflationists] also do not rely on 'interesting' or 'intended' models for languages that contain a truth predicate." In other words, deflationists about truth often do not want to commit to a standard model. No matter what one's philosophical feelings towards a deflationary position on truth, it would be unfaithful to rule it out of consideration before exploration has begun.

This desideratum leads to consideration of theories of truth over nonstandard syntaxes. A pre-theoretical thesis about truth would be that a theory of truth does not rule between, or decide upon, what our model of syntax is. We can perform many syntactic operations perfectly comfortably (if technically) within PA, and thanks to the completeness of first order logic, these syntactic operations apply within all models of the theory, including the nonstandard ones. It would certainly be a bold metatheoretic assumption to suggest that the truth predicate is privileged and should not be utilised within any nonstandard model. For the deflationist in

[4] In Chapter 3 I will aim to make these claims more precise, in order to understand what it means to be a deflationary theory of truth.

[5] Trying to clarify what these claims entail formally will be the aim of Chapter 4.

particular, who argues that truth lacks in 'causal-explanatory power', it would be highly undesirable if truth has so much power that it could rule out all other models of syntax alone. We should at least consider theories of truth over these models of syntax and deem them applicable to nonstandard sentences.

Even if one disagrees with this analysis, from a sceptical perspective, it is not entirely clear that we can even rule out nonstandard models of syntax. A nonstandard model appears a valid model of syntax which internally we could not distinguish, to our knowledge, from a standard model. This is, in the words of Dean (2013, p. 144), model-theoretic scepticism: "given that there are many non-isomorphic models of PA ... how is it that our arithmetical vocabulary latches on to 'the' standard model forumla?"

It seems impossible to conclusively argue that we are in an arithmetical world isomorphic to the standard model, rather than a nonstandard model, based solely upon our background theory of arithmetic. It appears that we learn the truth of arithmetical statements by deriving them deductively, implicitly within a formal system. This certainly at least matches with usual mathematical practice. The natural formal system to work within for these deductions is Peano Arithmetic, for example Isaacson (1987, p. 209) has argued that this is precisely what we mean by arithmetic. This formal system does not distinguish between the various arithmetical models, and thus, given our knowledge of arithmetic is derived from PA, we should be neutral on the question of which arithmetical model we should believe is 'real'. A sceptic should hold that she does not know there are not nonstandard numbers, and therefore a treatment of arithmetical truths must be equipped to deal with nonstandard sentences as well.

As sketched, this position of neutrality based only on PA does not sound wholly convincing, since we might argue our knowledge of arithmetic derives from further reflection (here, I use the word both informally and formally!) as well. For example, perhaps the theorist should believe she is in a world where the Gödel sentence is true, among other things. She might want to only consider models of arithmetic that believe these further sentences, and consider a stronger arithmetical theory than PA. Whilst I do not want to make such commitments and will take PA as my background theory, even the theorist who does shall still have multiple arithmetical models that she cannot formally distinguish whilst remaining in a first order arithmetical setting.

Perhaps the theorist we are considering wants to move to a stronger setting

than arithmetic altogether, however, such as ZFC. Models of this theory contain only one arithmetical world (ω), as one can prove full second order arithmetic is a consequence of ZFC. This means that one can prove Dedekind's theorem on the uniqueness of $\mathbb{N}$ within ZFC. This does not answer the sceptic's challenge, however, for her model of ZFC, V, could be an ω-nonstandard model of set theory. This is a model of set theory which does contain a unique arithmetical world (ω), but this is meta-theoretically considered a nonstandard model of arithmetic (Hamkins and Yang, 2013, p. 5-6). Even the move to a rich set-theoretic setting does not exclude nonstandard numbers from consideration. Such scepticism extends to computational defences of our knowledge of the standard world using Tennenbaum's Theorem as well, such as that provided by Halbach and Horsten (2005). Dean (2013), to hastily summarise a thorough and careful argument, analyses that this understanding of computation is also relative to a standard interpretation of computability and only rules out nonstandard models with the assumption that they are nonstandard from our perspective. From a sceptical position it appears very hard, if not impossible altogether, to rule out that our model of the world is 'nonstandard'.

Even if such scepticism appears unwarranted, this certainly ties back to my desideratum not to endorse too many metatheoretic commitments. To claim that we know we are within the standard model is to make a strong theoretical claim. This goes well beyond the formal commitments of second-order arithmetic and the usual set-theoretic and recursive framework of mathematics, which can be interpreted in nonstandard ways. In the exploration of minimal adequacy conditions for a theory of truth, it would be a strong addition to rule-out the sceptic's challenge immediately.

Scepticism and minimality aside, even the most ardent defender of our arithmetical/syntactic world as the standard model must admit that nonstandard models, numbers and sentences have mathematical use. Nonstandard analysis is known as often able to elegantly prove theorems of analysis that are otherwise complicated and indirect, for instance nonstandard proofs can provide simplification (Kanovei and Reeken, 1998) and construction (Leibman, 2005). Nonstandard arithmetic has similar utility. Many examples of this can be found in *Models of Peano Arithmetic* (Kaye, 1991), but the main use of nonstandard models is that the overspill principle allows stylish nonstandard proofs of theorems of $\mathbb{N}$ and that nonstandard numbers are able to code sets of infinitely many natural numbers. For a specific example,

Robinson and Roquette (1975) have demonstrated the usefulness of nonstandard arithmetic for proving results about diophantine equations. Nonstandard arithmetic is also a useful tool for reverse mathematics and exploration of the strength of various arithmetical theories, as Chong et al. (2014) show. There is mathematical interest in what these nonstandard models think of as true and false sentences, and thus these sentences should be considered as important to a theory of truth over these models.

Lastly, providing truth values for nonstandard sentences, within a theory, provides expressive utility to the theory. Some nonstandard sentences can be thought of as 'coding' infinitely-many standard sentences. For example, consider the sentences defined by:

$$\varphi_{n+1} = (\varphi_n \wedge \varphi_n)$$

where φ_0 is a standard sentence of arithmetic. Then φ_a, where a is a nonstandard number contains φ_n for every $n \in \mathbb{N}$. If a theory of truth says that φ_a is true, then this entails that infinitely many sentences are true - φ_n for every $n \in \mathbb{N}$. In a sense, it codes any number of conjunctions of φ_n, but within a single sentence. An account of truth for nonstandard sentences is useful, because it provides an account of truth for infinitely many standard sentences. Even the theorist who, against my claims above, is not interested in nonstandard arithmetic for its own sake, should be interested in what the nonstandard sentences can say about the standard sentences.

2.2.3 Deciding the Truth Values of Nonstandard Sentences

Given that having an account of truth which is accurate for nonstandard sentences is useful, there is still a barrier to overcome. Are we actually able to, externally, evaluate what truth values sentences of nonstandard length have? One may believe that having an accurate account of truth for nonstandard sentences is useful, but may also believe that we are not able to judge whether this account is actually accurate, since these sentences cannot be evaluated by us.

The nonstandard sentences are only nonstandard in the sense of their length. These sentences use the same logical connectives and quantifiers as the logician is used to, but can use infinitely many of them, in a non-wellfounded sequence. Nonstandard sentences can also contain nonstandard terms: terms built in the same was as standard terms, but by iterating functions infinitely many times in

a non-wellfounded way. I now argue that this non-wellfounded length and use of infinitely-many connectives, quantifiers or functions is unproblematic, we can still identify their truth values. There are many sentences in and about logic, and even used by philosophers of truth, which are deemed perfectly valid, well-understood and true, despite being, infinitely long. I will argue that we can judge the truth values of these sentences and, from this, assign truth values to some nonstandard sentences as well.

Firstly, I would like to distinguish here between an expression of a sentence and the sentence itself. The sentence being a well-formed syntactic object, and the expression being a specific mention (oral, written or otherwise) of this object. It is true that I do not have any examples of infinitely long expressions in use. This section would certainly be tedious to read were I to produce an example of one! What we do have, however, are expressions which stand in for sentences which are infinitely long.

A good example of this usage of infinitely long sentences within formal logic is in looking at axiom and theorem schemata. These are infinitely long lists of axioms or theorems (usually) for every formula definable within the formal language. An example of this is the axiom of induction scheme in Peano Arithmetic, which this chapter is working within. This is specifically formulated as an infinite list, rather than a finite variant, because it is not possible to finitely axiomatise induction in a first order way (Kaye, 1991, p. 132). An infinite list of sentences is a common appearance in mathematical logic and, by taking the conjunction of all of these sentences (now working in our metatheory), one can easily produce an infinitely long sentence about formal logic.

A critic may propose that this list is not actually infinitely long, but merely contains as many instances of the schema as required. For example, one can only consider finitely many proofs, all of which use finitely many axioms, and thus only finitely many instances of the axiom schema are ever used and the list can be considered as finitely long. This does not match up with uses of the axiom schema, however. Certain model-theoretic constructions involve enumerating all of the axioms and would not work with only a finite subset of the axioms. Therefore, this must be considered as a genuinely infinite list for it to match up with common mathematical practice.

A more ironic example of this behaviour comes from the definition of finite in first order logic. Even to express that a given domain M is finite one must use a

countable list of sentences. It is a theorem of first order logic that the sentence "M has infinitely many elements" is not finitely statable (Kaye, 2007, p. 146) and thus the sentence "M has finitely many elements" is not either.

One is able to express that a given domain M is finite, however. This can be formed by taking the infinitely long disjunction:

$$\begin{gathered}\text{“there is one object”} \vee \text{“there are two objects”} \vee \\ \text{“there are three objects”} \vee \text{ “there are four objects”} \vee \ldots\end{gathered} \tag{∇}$$

It is important to note that this sentence, whilst it is a sentence of the meta-theory, can also be a genuine sentence of a formal language. If we work within the language of $\mathscr{L}_{\omega_1\omega}$, a formal language where conjunctions and disjunctions are allowed to appear countably many times in a sentence, then this is a perfectly acceptable formal sentence of infinite length. There is no barrier to infinitely long sentences built into logic and some logics such as ω-logic make powerful use of them.

This shows that infinitely long sentences are not anathema to logic, but instead a common element of it. We regularly judge the truth values of these sentences, and thus a sentence of infinite length should be deemed accessible.

In fact, looking beyond formal logic, one can find infinitely long sentences in use. Philosophers of truth have used infinitely long sentences in their own writing. An example of this is that Quine (1986, p. 12) and Horwich (1998, p. 4), among others, have stated that the sentence "every proposition is either true or false" is nothing more than the infinite conjunction

$$\begin{gathered}\langle p_1\rangle \text{ is either true or false } \textbf{and } \langle p_2\rangle \text{ is either true or false } \textbf{and} \\ \langle p_3\rangle \text{ is either true or false } \textbf{and} \ldots\end{gathered} \tag{χ}$$

where $p_1, p_2, \ldots$ is an infinite list of all propositions and $\langle p_i\rangle$ is an expression of p_i.

This example shows that infinitely long sentences are acceptable sentences which can be examined and assigned truth values, even according to philosophers of truth. A sentence with infinite length is dealt with perfectly adequately by our standard reasoning and mathematical practice.

I argue that we can move from the truth values of sentences of infinite length to the truth values of sentences of nonstandard length. To demonstrate this, I will

positively argue for the falsity of

$$(0 = 1 \vee (0 = 1 \vee (0 = 1 \vee (\ldots \vee (0 = 1 \vee 0 = 1) \ldots)))) \qquad (\dagger_a)$$

where there are a disjuncts $0 = 1$, where $a \in M \backslash \mathbb{N}$ and M is a nonstandard model of arithmetic.

The sentence

$$(0 = 1 \vee (0 = 1 \vee (0 = 1 \vee (0 = 1 \vee (\ldots))))) \qquad (\text{II})$$

is similar to the above examples ∇ and χ, in so far as its comprehensibility, and is definitely false. The above sentence II represents the disjunction of countably many sentences $0 = 1$. It has a clear structure which is analogous to the natural number line, where every odd number is the sentence $0 = 1$ and every even number is a disjunction symbol. Clearly, this is an infinitely long sentence.

The sentence

$$(0 = 1 \vee (0 = 1 \vee (0 = 1 \vee (\ldots \vee (0 = 1 \vee 0 = 1) \ldots)))) \qquad (\dagger_a)$$

where there are $a \in M \backslash \mathbb{N}$ disjuncts $0 = 1$, has a trickier structure, but has the same truth value as II. One can take II and re-order the disjuncts, which will turn the sentence into $\dagger_a$, without affecting its truth value. Whilst II has structure analogous to $\mathbb{N}$, the sentence $\dagger_a$ has a more complicated structure, which is analogous to $\mathbb{N} \cup (\{z \in \mathbb{Z} : z < n\} \times \mathbb{Q})$ for some $n \in \mathbb{N}$. Since both of these structures are countable, there is a function which maps each structure onto the other. This function will take the sentence II and reorder the structure of its disjuncts into $\dagger_a$. This reordering is not unusual mathematical practice, even for infinitely long sequences, and it is a common theorem that reordering a sequence of disjuncts does not affect the truth-value of the sentence. Given that II is false, one applies the common mathematical technique of reordering its disjuncts, then applies the theorem that a reordering of disjuncts does not affect a sentence's truth value, to arrive at the conclusion that $\dagger_a$ is definitely false.

Therefore there is a sentence of nonstandard length that one can analyse the truth value of and accept as a reasonable and meaningful sentence. It would therefore be desirable for a theory of truth over a nonstandard model of arithmetic containing a to decide that this sentence is indeed false. Surprisingly, this is a

nontrivial matter however, and one which I will now explore further.

2.3 Pathologies and Compositional Truth

2.3.1 Compositional Truth as a Minimum

I shall now explore in detail how minimal theories of truth behave over nonstandard syntaxes. To do this I shall be looking at models of compositional truth (CT^-) as a minimally adequate semantic interpretation of the truth predicate. In what follows I shall often identify a formula with its Gödel code for ease of reading and shall make use of the dot notation detailed in Section 1.2 in specifying the axioms of CT^-.

Definition 2.3.1.1. *The axioms of* CT^- *are the axioms of* PA *and the following:*

$\mathrm{CT1}: \forall m, n[\mathrm{ClTerm}(m) \wedge \mathrm{ClTerm}(n) \rightarrow (\mathrm{Tr}(m \doteq n) \leftrightarrow \mathrm{Val}(n) = \mathrm{Val}(m))]$

$\mathrm{CT2}: \forall m, n[\mathrm{ClTerm}(m) \wedge \mathrm{ClTerm}(n) \rightarrow (\mathrm{Tr}(m \dot{<} n) \leftrightarrow \mathrm{Val}(n) < \mathrm{Val}(m))]$

$\mathrm{CT3}: \forall \alpha, \beta[\mathrm{Sent}(\alpha \dot{\wedge} \beta) \rightarrow (\mathrm{Tr}(\alpha \dot{\wedge} \beta) \leftrightarrow \mathrm{Tr}(\alpha) \wedge \mathrm{Tr}(\beta)]$

$\mathrm{CT4}: \forall \alpha, \beta[\mathrm{Sent}(\alpha \dot{\vee} \beta) \rightarrow (\mathrm{Tr}(\alpha \dot{\vee} \beta) \leftrightarrow \mathrm{Tr}(\alpha) \vee \mathrm{Tr}(\beta)]$

$\mathrm{CT5}: \forall \varphi[\mathrm{Sent}(\varphi) \rightarrow \mathrm{Tr}(\dot{\neg} \varphi) \leftrightarrow \neg \mathrm{Tr}(\varphi)]$

$\mathrm{CT6}: \forall \varphi \forall x[\mathrm{Sent}(\dot{\exists} x \varphi) \rightarrow (\mathrm{Tr}(\dot{\exists} x \varphi) \leftrightarrow \exists b \mathrm{Tr}(\varphi(\dot{b}))]$

$\mathrm{CT7}: \forall \varphi \forall x[\mathrm{Sent}(\dot{\forall} x \varphi) \rightarrow (\mathrm{Tr}(\dot{\forall} x \varphi) \leftrightarrow \forall b \mathrm{Tr}(\varphi(\dot{b}))]$

Let $M \vDash \mathrm{PA}$. *A set* $S \subseteq M$ *is a satisfaction class*[6] *for* M *if and only if* $(M, S) \vDash \mathrm{CT}^-$.

A satisfaction class is therefore a set of codes of sentences from the model where the sentences satisfy the compositional properties of truth. This is a typed theory of truth, so the truth predicate only ranges over formulas of the base language (in my case, arithmetic) and thus does not apply to itself. We do not specify any

[6] A brief note on terminology: I follow Halbach (2011, Def. 8.14) in defining satisfaction classes as models of CT^-, rather than treating these as models of a compositional satisfaction theory as is usually done in the literature. This is not a problem formally, since I use a numeral variant of CT^-. For the details see Cieśliński (2017, p. 110).

induction axioms for the language with the truth predicate, only for formulas of the base language.

This is, in my mind, an absolute minimum of how we want a theory of truth to behave. It should be able to talk of the truth or falsity of all sentences of the base language and satisfy the familiar compositional properties that are widely believed for these sentences. Further, it entails the T-schema for all sentences of standard length, and thus is sufficiently strong to perform a basic role of truth.[7] If a theory of truth did not do this, then it would not be sufficient to perform any of the truth predicate's role in language.

There are further properties of truth that one may want, such as the ability to speak of the truth and falsity of sentences containing the truth predicate, among other things. I will not build this into my theory of truth as an absolute minimum. In developing a theory of truth for arithmetic, it is arithmetical sentences that are of primary interest, not sentences about arithmetic and truth. Discussing the drawbacks and benefits of various type-free approaches to truth would no longer be about minimal considerations. Further, even for the theorist who disagrees and does want a stronger baseline for truth, the details of what follows will still apply to her. She will still want these minimal considerations about truth within her theory, regardless of her other concerns.

The following theorems are useful theorems from the study of satisfaction classes.

Theorem 2.3.1.2 (Lachlan's Theorem)**.** *If $M \vDash \mathrm{PA}$, M is nonstandard and M has a satisfaction class S, then M is recursively saturated (Kotlarski, 1991, Theorem 3).*

Theorem 2.3.1.3 (KKL's Theorem)**.** *If $M \vDash \mathrm{PA}$ and M is countable, nonstandard and recursively saturated, then M has a satisfaction class S (Kotlarski et al., 1981, Main Theorem).*

Theorem 2.3.1.4. *If $M \vDash \mathrm{PA}$, M is countable and nonstandard and M has a satisfaction class S, then M has $2^{\aleph_0}$-many such satisfaction classes (Kotlarski, 1991, Theorem 1).*

[7] One may suggest that the T-schema alone would be better as a minimal condition upon truth. This is not able to, formally, prove generalisations of the compositional clauses, however (Halbach, 2011, Theorem 7.6). A theory of truth should prove these, in order to be of sufficient strength. By using the compositional clauses as the axioms of the theory one gets around this issue immediately.

Theorem 2.3.1.5. *If $M \vDash \mathrm{PA}$, σ is a (standard) $\mathscr{L}_A$ sentence and M has a satisfaction class S, then $(M, S) \vDash S(\ulcorner \sigma \urcorner)$ if and only if $M \vDash \sigma$ (Halbach, 2011, Lemma 8.4).*

2.3.2 Pathological Satisfaction Classes

I shall now examine a weakness of the theory of satisfaction classes. Theorem 2.3.1.4 says that a countable nonstandard model of arithmetic possesses $2^{\aleph_0}$ satisfaction classes (if it possesses any), and thus different satisfaction classes give different truth values for some sentences of arithmetic. Since, by Theorem 2.3.1.5 they all agree on the standard sentences, these must be nonstandard sentences of arithmetic. This disagreement can be highly counter-intuitive, in that some of these sentences appear always false from an external metatheoretic perspective. These sentences are known as pathologies and the satisfaction classes which exhibit this behaviour as pathological satisfaction classes.

In order to examine this, we must first fix a nonstandard model M of Peano Arithmetic. This model is chosen to be countable and recursively saturated and thus, by Theorem 2.3.1.3, has a satisfaction class. We also fix a nonstandard number $a \in M$.

Kotlarski et al. (1981) provide a criterion of which sentences can be true in a satisfaction class. This theorem entails the existence of pathological sentences. To state this I first introduce the notion of an approximation of a formula.

Definition 2.3.2.1. *Let $\mathscr{L}'_A(M) = \mathscr{L}_A(M) \cup \{P_i : i \in \mathbb{N}\}$ where each P_i is a predicate symbol. In particular we note that all sentences of this language are of standard length. For any sentence $\psi[P_0, P_1, \ldots, P_n]$ of $\mathscr{L}'_A(M)$ we denote by $\psi[{}^{\pi_0}/_{P_0}, {}^{\pi_1}/_{P_1}, \ldots, {}^{\pi_n}/_{P_n}]$ the formula which results from replacing all predicate symbols P_i by formulas $\pi_i(\bar{x}_i)$ of the language $\mathscr{L}'_A(M)$ or ${}^*\mathscr{L}_A(M)$.*
Let φ be a formula of $\mathscr{L}'_A(M) \cup {}^*\mathscr{L}_A(M)$. If it is possible to write φ as some formula $\psi[{}^{\pi_0}/_{P_0}, {}^{\pi_1}/_{P_1}, \ldots, {}^{\pi_n}/_{P_n}]$ as above, then we say that $\psi[{}^{\pi_i}/_{P_i}]$ is an approximation of φ. (Kotlarski et al., 1981, p. 286)

An example of this can be provided. Consider the sentence

$$(\sigma \vee (\sigma \vee (\sigma \vee (\ldots \vee (\sigma \vee \sigma) \ldots)))) \qquad (\varphi)$$

where there are a connectives $\vee$, σ is a standard $\mathscr{L}_A$ sentence and $a \in M$ is

nonstandard. This can be approximated by a new formula. We can write φ as

$$(\sigma \vee (\sigma \vee (\sigma \vee (\ldots \vee (\sigma \vee P)\ldots))))$$

where there are only $n \in \mathbb{N}$ connectives $\vee$ and P is some predicate symbol which stands for the rest of the sentence. This can be substitued by the nonstandard formula π to get the new formula

$$(\sigma \vee (\sigma \vee (\sigma \vee (\ldots \vee (\sigma \vee \pi)\ldots))))$$

which is an approximation of the original sentence.

This definition of an approximation of a sentence is crucial to the study of pathologies, as it provides an exact characterisation of which sentences can be true in a satisfaction class, even if they ought not to be.

Theorem 2.3.2.2. *Let M be a countable, nonstandard and recursively saturated model of Peano Arithmetic and φ a sentence of ${}^*\mathscr{L}_A(M)$. There exists a satisfaction class S which contains φ if and only if there is no approximation ψ of φ such that $Th(M) \vdash \neg\psi$ (Kotlarski et al., 1981, p. 292).*

For example, this theorem implies that the previous example φ will always be contained in some satisfaction class. It was shown that

$$(\sigma \vee (\sigma \vee (\sigma \vee (\ldots \vee (\sigma \vee \pi)\ldots))))$$

is an approximation of φ. In fact, the only approximations of φ are of this form, with $n \in \mathbb{N}$ disjuncts $\vee$, since the approximation must be of standard length.

Using standard first order logic, one will never be able to prove with $Th(M)$ that any of these approximations are definitely false, since the formula π might be true, which would make the approximation true.

That this theorem provides unintuitive ('pathological') satisfaction classes can be seen immediately from this. If σ is a false sentence, then the sentence φ is intuitively false as well. This theorem also provides many more ways of producing pathological sentences, however, which can be seen in the following examples given in the literature.

There are many examples of pathological sentences that follow the example above. Cieśliński (2010b, p. 327) provides the simple example of a sentence that

can be made true by a satisfaction class, by following the exact reasoning above. This is the sentence:

$$\delta_a^{(0\neq 0)}, \text{ where}$$
$$\delta_0^{(0\neq 0)} \text{ is } (0 \neq 0) \text{ and}$$
$$\delta_{n+1}^{(0\neq 0)} \text{ is } (\delta_n^{(0\neq 0)} \vee \delta_n^{(0\neq 0)}) \text{ for all } n \in M.$$

There are alternative classes of examples that can be produced, however. Engström (2002, p. 56) provides a number of these different examples of pathological statements. The first[8] of these is the sentence:

$$\exists x_0, x_1, ..., x_a[0 \neq 0] \qquad (\ddagger_a^{(0\neq 0)}))$$

He also provides a number of different pathological types of sentences which are relative to some sentence φ in $\mathscr{L}_A$. These are the sentences:

$$(\exists x_0, x_1, ..., x_a[\varphi]) \leftrightarrow \neg\varphi \qquad (\Diamond_a^{\varphi})$$

$$(\exists x_0 \forall x_1 \exists x_2 ... \forall x_{2a-1} \exists x_{2a}[\varphi] \leftrightarrow \neg\varphi, \text{ and} \qquad (\heartsuit_a^{\varphi})$$

$$\epsilon_a^{\varphi}, \text{ where}$$
$$\epsilon_0^{\varphi} \text{ is } \neg(\varphi \vee \neg\varphi) \text{ and}$$
$$\epsilon_{n+1}^{\varphi} \text{ is } \epsilon_n \vee \epsilon_n \text{ for all } n \in M.$$

All of the above sentences can be believed to be true by a satisfaction class for M, but by common intuition are not true. Further, it appears obvious from our metatheoretic perspective that no good truth definition should view them as true. This is why they are considered as pathologies.

The reason that Theorem 2.3.2.2 applies to these sentences and they can be made true by a satisfaction class is because the theory can only 'examine' finitely many connectives and quantifiers, because there is no induction specified for the language with the truth predicate. If a nonstandard number of connectives or quantifiers appear in a sentence σ, then only finitely many parts of the sentence are examined in an approximation of it and, if this is consistent with what is

[8]This is not actually the first pathological statement he considers, the first is that certain nonstandard terms can be equal to (the 'wrong') constant symbols for a nonstandard number. This does does not arise in my context due to the way that truth for atomic formulas is defined here, which differs to Engström's.

already known, then it can be believed true by a satisfaction class.

2.3.3 Alternative Pathological Sentences

The examples given above are the typical pathological sentences that can be found in the literature. There are other pathological sentences that can be considered, however, and I will provide some examples of these. These sentences are useful as they provide further understanding of what the pathological sentences are, as well as providing some interesting examples which go beyond the current literature.

It has been widely remarked that a nonstandard disjunction of single false sentence can be true in a satisfaction class, but the problem is deeper than this. Any nonstandard disjunction of sentences can be true in a satisfaction class, even if all the disjuncts are different. See, for example:

$$(0 = 1 \vee (0 = \underline{2} \vee (0 = \underline{3} \vee (\ldots \vee (0 = \underline{a-1} \vee 0 = \underline{a}) \ldots)))) \qquad (\ominus_a)$$

When one of these disjuncts is true, then it is clear that the sentence is not pathological, in the sense that the entire sentence should be true. When all of these disjuncts are false, however, then the sentence is pathological. I will denote sentences where all disjuncts are the identical sentence φ by δ_a^φ.

Another type of pathological sentence, which to the author's knowledge has not appeared in the literature, is one involving a nonstandard number of negation symbols. For fixed $\varphi \in {}^*\mathscr{L}_A(M)$ this is the sentence:

$$\mho_a^\varphi, \text{ where}$$
$$\mho_0^\varphi \text{ is } \varphi, \text{ and}$$
$$\mho_{n+1}^\varphi \text{ is } \neg\mho_n^\varphi \text{ for all } n \in M.$$

For φ such as $(0 = 1)$, this sentence should be intuitively false when a is even and true when a is odd, but a satisfaction class can believe that $\mho_{2a}^{(0=1)}$ is true and $\mho_{2a+1}^{(0=1)}$ is false. In fact, the truth-value of one, entails the truth value of the other through the compositional clause for negation.

This example shows that pathological sentences can be sentences which are intuitively true, but are not always believed to be true by a satisfaction class. Another example, for a given $\varphi \in {}^*\mathscr{L}_A(M)$, of this type is the sentence:

$$*_a^\varphi, \text{ where}$$
$$*_0^\varphi \text{ is } (\varphi \wedge \varphi), \text{ and}$$
$$*_{n+1}^\varphi \text{ is } (*_n^\varphi \wedge *_n^\varphi) \text{ for each } n \in M.$$

The sentence $*_a^{(0=0)}$ is then intuitively true, but can be believed to be false by a satisfaction class.

The family of sentences which are intuitively true, but can be false according to satisfaction classes, can be considered as the dual of the previous examples. If a sentence is intuitively true, then its negation is intuitively false, and if it can be false in a satisfaction class, then its negation can be true in a satisfaction class, by the compositional clause for negation. This means that a treatment of the intuitively false pathologies will result in a treatment of the intuitively true pathologies as well, and thus I shall only consider the intuitively false pathologies from here on.

One last remark of note, is that, due to the nature of the compositional axioms for truth, if a satisfaction class contains one pathological sentence, then it will contain many. For example, if φ is a pathological sentence in a satisfaction class S, then $\neg\neg\varphi$, $\varphi \vee \varphi$, $\varphi \wedge \varphi$, etc. will also be in S. This is because the definition of a satisfaction class forces this to be the case and shows that the existence of just one pathology generates a family of controversial instances.

2.4 Robinson Semantics

It was shown in the previous section that satisfaction classes, minimal interpretations of a truth predicate, can contain 'pathological' sentences of arithmetic. There is a wide spread of examples of these sentences which are intuitively false, and which mathematical reasoning says must be so, and thus should not be contained in an interpretation of the truth predicate. The theory of satisfaction classes, CT^-, is not sufficiently strong to prove that they are false.

Capturing the notions of 'intuitively false' and 'mathematical reasoning' that are playing this role is no easy task, of course. Ordinary external semantic definitions of Truth and Falsity (here, I shall capitalise the words True and False to denote metatheoretic external semantic values, rather than those given by the internal truth predicate that is being considered) only apply to sentences of stand-

ard length. The usual model-theoretic definition of Truth does not apply to these nonstandard sentences which are intuitively False. To say that these sentences are False requires an extended external semantic definition of Truth and Falsity. Robinson (1963, p. 106-7) provides the beginnings of such a definition. In what follows I shall adapt this definition to be as wide as possible, in order to capture as many pathologies as possible. I shall denote this with the consequent relation $\models^*$.

Robinson's definition applies only to those formulas which are simple (built from standard terms) and containing only finitely-many alternating connectives. This allows one to assign a value of True or False to many sentences of ${}^*\mathscr{L}_A(M)$ and can be used to generate an extended T-schema. This leaves more complicated (but still accessible) sentences completely unfixed, however. I shall introduce an expanded definition, inspired by Robinson, which provides relational truth-values across the entirety of ${}^*\mathscr{L}_A(M)$. This does not fix every formula as True or False, but does fix a coherent structure between the formulas.

Definition 2.4.1. *For each φ of ${}^*\mathscr{L}_A(M)$ we provide the partial definition that $M \models^* \varphi$ if one of the following conditions hold:*

- *If φ is atomic, then M computes the values of terms and constants in φ in accordance with what φ states.*[9]
- *$M \models^* \neg\varphi$ if and only if $M \not\models^* \varphi$.*
- *$M \models^* \bigwedge_{i<a} \varphi_i$ if and only if $M \models^* \varphi_i$ for all i from 1 to a.*
- *$M \models^* \bigvee_{i<a} \varphi_i$ if and only if $M \models^* \varphi_i$ for some i from 1 to a.*
- *For a string of quantifiers Q, we have $M \models^* Q\varphi$ if and only if $M \models^* \xi(x_1, x_2, ..., x_p, f_1(\bar{y_1}), f_2(\bar{y_2}), ..., f_q(\bar{y_r}))$ for all possible substitutions. Here, $\forall x_1 \forall x_2 \ldots \forall x_p \xi(x_1, x_2, \ldots, x_p, f_1(\bar{y_1}), f_2(\bar{y_2}), ..., f_q(\bar{y_r}))$ is the Skolemised form of $Q\varphi$, where each $f_i(\bar{y_j})$ is one of the Skolem functions.*

Our definition above, as Robinson shows, defines the Truth or Falsity of many sentences of nonstandard complexity. For 'right-bracketed' sentences involving only a finite number of alternating connectives or quantifiers, we have a fixed

[9]My thanks and acknowledgements to an anonymous referee of an article based on this chapter for the suggestion that atomic formulas can be defined in this way.

external valuation of whether they are True or False. For sentences containing a nonstandard number of alternating connectives or quantifiers, we do not get a fixed valuation, but know their relational truth value in terms of other nonstandard sentences. This gives us a formal explanation of the reasoning that many pathologies considered above, based only on finitely alternating connectives, are False. For example, consider the sentence:

$$\exists x_0 \forall x_1 \exists x_2 ... \forall x_{2a-1} \exists x_{2a} [\dagger_b]$$

Where $\dagger_b$ is the sentence:

$$(0 = 1 \vee (0 = 1 \vee (0 = 1 \vee (. . . ... \vee (0 = 1 \vee 0 = 1) . . .))))$$

For $b \in M \backslash \mathbb{N}$ disjuncts, this sentence is a pathology, since it can be true in a satisfaction class, but should intuitively be false. This sentence is made False by the definition of $\vDash^*$ above. If $M \vDash \mathrm{PA}$ is nonstandard and $a, b \in M \backslash \mathbb{N}$ we get that $M \nvDash^* 0 = 1$. Then, we get that $M \nvDash^* \dagger_b$ and thus $M \nvDash^* \exists x_0 \forall x_1 \exists x_2 ... \forall x_{2a-1} \exists x_{2a} [\dagger_b]$.

This definition cannot fix valuations for all the pathologies, however, and certainly those not containing a nonstandard number of alternating connectives or quantifiers. For example, consider the family of formulas defined as follows: ζ_0 is $(0 \neq 0)$ and ζ_{i+1} is $\exists x_i (\zeta_i \vee \zeta_i)$. These sentences alternate between the existential quantifier and disjunction and our definition of $\vDash^*$ cannot finitely evaluate this formula for nonstandard i. Applying our rules above we learn $M \vDash^* \zeta_i$ if and only if $M \vDash^* \zeta_{i-1}$ if and only if $M \vDash^* \zeta_{i-2}$, etc. This sequence is not well-founded, and so never terminates. What it does at least offer, however, are relational alethic values. We may not be able to state that $M \nvDash \zeta_i$, but we do gain the information that $M \nvDash \zeta_i$ if and only if $M \nvDash \zeta_{i-1}$ and $M \nvDash \zeta_{i+1}$.

Another weakness of Definition 2.4.1 is that it only applies to conjunctions or disjunctions which make use of 'right-bracketing'. As a quirk of my definition of the shorthand $\bigwedge_{i<n} \varphi_i$ and $\bigvee_{i<n} \varphi_i$ in Section 1.2, very similar, but alternatively bracketed, sentences are not spoken for. For example, the sentence:

$$((((. . . (0 = 1 \vee 0 = 1) \vee . . .) \vee 0 = 1) \vee 0 = 1) \vee 0 = 1)$$

with $b \in M \backslash \mathbb{N}$ disjuncts is extremely similar to $\dagger_b$ above, but with alternative bracketing. This is technically a different pathology and Definition 2.4.1 cannot

say that it is False. It is hard to easily avoid this issue for all the many differently-bracketed, but similar, sentences, without arbitrarily restricting what counts as a well-formed formula. Finding a method for doing this remains an open question.

Question 2.4.2. *Is there a natural way to extend Definition 2.4.1 to account for sentences containing a repeated nonstandard disjunction or conjunction which are formed with 'alternative' (not-right) bracketing?*

Despite these downsides, however, Definition 2.4.1 is still an extension of the standard semantics for True and False into the nonstandard domain, and one which can deal with many pathologies. Taking $\vDash^*$ for only 'right-bracketed' finitely-alternating connectives or quantifiers offers a mathematically precise and formal method of stating that many pathologies are False. It gives external semantic valuation of nonstandard sentences, as I argued in Section 2.2.3 was indeed possible for some nonstandard sentences. Whilst it does not offer this for sentences containing a nonstandard number of alternating connectives or quantifiers, or those using alternative bracketing, it does at least provide external relations between the semantic values for these. These relations are intuitively plausible and even if we cannot state whether the formula is True or False, it does match intuitions that we know whether these formulas are True or False in relation to other formulas. It appears to be a natural minimal condition for a theory of truth to satisfy. This leads to the question of how a truth predicate should react to these new semantic valuations, which I shall now explore.

2.5 An Extended T-Schema

I propose that the notion of semantic entailment $\vDash^*$ can be thought of as providing an extended T-schema. It provides a T-schema for sentences belonging to a nonstandard model of syntax and should be taken as a new minimal condition for a semantic theory of truth considered over nonstandard syntaxes.

The T-schema is presented within natural language as the schema that $\ulcorner p \urcorner$ is true if and only if p, where $\ulcorner p \urcorner$ ranges over all truth-bearers considered. This is interpreted model-theoretically in the following manner:

Definition 2.5.1 (T-Schema)**.** *An instance of the T-schema (TS), for* $M \vDash \mathrm{PA}$*,*

has the following form:

$$M \vDash \varphi \textit{ if and only if } (M, Tr) \vDash Tr(\ulcorner \varphi \urcorner)$$

where Tr is an interpretation of the truth predicate and $\ulcorner \varphi \urcorner$ ranges over all sentences of the language.

This is certainly a valid interpretation of ES for natural language, but suffers from a lack of applicability to all sentences inherent within the intended schema. For the definition to make sense, it implicitly assumes that the language is the standard interpretation of the language, and not a nonstandard syntax instead. If φ is a nonstandard sentence, then the standard semantic notion $\vDash$ of True and False will not apply. This is a gap when M is a nonstandard model, since there will be cases where $(M, Tr) \vDash Tr(a)$ and $M \vDash Sent(a)$, but the biconditional cannot be formed, as the sentence that a is the code of is nonstandard. This is highly undesirable when considering nonstandard syntaxes and nonstandard models, as I argued in Section 2.2.2 that we ought.

The pathological examples considered in Section 2.3.2 provide good examples of why we ought to reject the T-schema as sufficient at capturing all the natural language T-schema aims to entail. In these pathologies we have sentences which are False, but that many semantic interpretations of a theory of truth say are true.

Consider the simple example of $\delta_a^{(0\neq 0)}$, where:

$$\delta_0^{(0\neq 0)} \text{ is } (0 \neq 0) \text{ and}$$
$$\delta_{n+1}^{(0\neq 0)} \text{ is } (\delta_n^{(0\neq 0)} \vee \delta_n^{(0\neq 0)}) \text{ for all } n \in M.$$

It is the case that $\neg\delta_a^{(0\neq 0)}$, but the corresponding biconditional for the theory of satisfaction classes does not hold. The T-schema, when interpreted as above, is not sufficiently strong to rule out undesirable theories of truth. It is insufficient as a minimum condition on theories of truth.

By utilising the stronger notion of semantic entailment, we can introduce an extended T-schema to avoid this issue. This extended T-schema avoids these worries and is a sufficient minimum condition for theories of truth. This has the following formulation:

Definition 2.5.2 (Extended T-Schema)**.** *An instance of the extended T-schema*

(ETS) has the following form:

$$M \vDash^* \varphi \textit{ if and only if } (M, Tr) \vDash Tr(\ulcorner \varphi \urcorner)$$

where Tr is a truth predicate and $\ulcorner \varphi \urcorner$ ranges over all sentences of the model M's interpretation of its language.

It is clear that when the model considered is the standard model, then EES collapses to ES. This is because the $\ulcorner \varphi \urcorner$ will only range over the standard sentences of the language, and so for any such $\ulcorner \varphi \urcorner$ it is the case that $M \vDash^* \varphi$ if and only if $M \vDash \varphi$.

When the model examined is a nonstandard model, EES becomes an interesting stronger condition than ES. The extended T-schema contains every instance of ES, since every standard sentence φ will still be considered as sentences within the model's interpretation of its language, but goes beyond it to provide detail on sentences of nonstandard length as well. EES is a useful extension of the T-schema for nonstandard models.

This is a natural condition to propose for a theory of truth. Given the reasoning provided in Section 2.2.3, one is able to assign truth values to some nonstandard sentences and thus it should be expected that a theory of truth will respect these. It would be highly undesirable for a theory of truth to state that a sentence is true, when it is actually False, or for it to state that a sentence is false, when it is actually True.

In fact, this can be thought of as the correct interpretation of the natural language T-schema for nonstandard models. The intention of the natural language schema is to capture that p holds within a model if and only if it is true. Usually, p holding within a model is denoted by $\vDash$, but as stated earlier this only applies when p belongs to a standard syntax. When p is of a nonstandard syntax, the natural language T-schema should still apply, but should instead be formulated by using a better definition of p holding within a model. We have a better definition of p holding within a model with $\vDash^*$.

This avoids issues of pathologies inherent within CT^- as well. As stated previously, it is the case for many pathologies φ, those with only finitely many alternating connectives or quantifiers, that $M \vDash^* \neg\varphi$. The extended T-schema rules out many pathological satisfaction classes as appropriate theories of truth. Further, the extended T-schema enforces additional structure within a satisfaction

class and ensures that externally plausible equivalences hold within. This includes equivalences such as $(M, Tr) \vDash Tr(\ulcorner \bigvee_{i<a} \varphi_i \urcorner)$ if and only if $(M, Tr) \vDash Tr(\ulcorner \varphi_i \urcorner)$ for some $i < a$. We can think of these as compositional clauses for a nonstandard syntax. The extended T-schema shows that the theory of satisfaction classes does not reach an appropriate minimum for how a theory of truth should behave. This is correct, for the pathologies are problematic and should be avoided. The extended T-schema is a good explication of this and makes up for the weaknesses within the usual T-schema.

2.5.1 The Extended T-Schema and CT$^-$

Having accepted that the extended T-schema is the correct way of thinking about the T-schema for nonstandard syntaxes, the natural question to ask is what are the effects of taking this as our schema? I shall again be asking this question from the position that CT$^-$ are the minimum axioms that we want truth to satisfy.

It turns out that, when EES is taken as a minimal adequacy condition for truth, then CT$^-$ is no longer a sufficient minimum. In fact, closing a satisfaction class under EES results in a stronger theory that is nonconservative over PA. It is able to prove all $\mathscr{L}_{Tr}$-consequences of CT$^-$ + IΔ_0, the theory of compositional axioms for truth and induction for all Δ_0 formulas in the language with the truth predicate. First, I shall introduce the notation CT$^-$+ $\vDash^*$ to denote a theory of a class of models.

Definition 2.5.1.1. *Denote by* CT$^-$+ $\vDash^*$ *the theory of the class of models* (M, Tr) *where* $M \vDash$ PA*,* Tr *is a satisfaction class for* M *and* $M \vDash^* \varphi$ *if and only if* $(M, \mathrm{Tr}) \vDash \mathrm{Tr}(\ulcorner \varphi \urcorner)$*.*

Note that a full induction scheme for $\mathscr{L}_{Tr}$, i.e. CT, suffices to imply CT$^-$+ $\vDash^*$. With induction we can prove that the Robinson inspired semantic consequence conditions hold within the satisfaction class. It is an open question what the exact strength of the theory CT$^-$+ $\vDash^*$ is, whether this is full CT or some weaker subtheory.

Question 2.5.1.2. *Is there a natural theory of truth* CT$^-$ + X*, for some* X*, which is equivalent to* CT$^-$+ $\vDash^*$*?*

Whilst it may not be known what the exact provability strength of CT$^-$+ $\vDash^*$ is, the theory, with its strengthened T-schema, can prove generalised compositional

clauses. These are generalised compositional clauses which express, informally, that the standard compositional clauses hold for any number of connectives. Two of these clauses: disjunctive correctness (DC) and conjunctive correctness (CC) are already well-known within the literature and express that a disjunction (conjunction, respectively) of a number of formulas is true if and only if at least one (all) of the formulas is (are) true. Łełyk (2017) provides a thorough discussion of their use and history. The final clause, quantifier correctness (QC), is new and expresses that a formula beginning with a number of quantifiers is true if and only if the Skolemised version of this formula is true for all variable substitutions. These compositional clauses are intuitively valid and are simply extensions of the usual clauses, expanding one connective to a number of connectives.

Theorem 2.5.1.3. *The theory* $\mathrm{CT}^{-}+\vDash^{*}$ *implies the following principles:*

- (DC) $\forall c[\mathit{SetSent}(c) \rightarrow (\mathrm{Tr}(\ulcorner\bigvee_{\varphi\in c}\varphi\urcorner) \leftrightarrow \exists\ulcorner\varphi\urcorner \in c\ \mathrm{Tr}(\ulcorner\varphi\urcorner))]$
- (CC) $\forall c[\mathit{SetSent}(c) \rightarrow (\mathrm{Tr}(\ulcorner\bigwedge_{\varphi\in c}\varphi\urcorner) \leftrightarrow \forall\ulcorner\varphi\urcorner \in c\ \mathrm{Tr}(\ulcorner\varphi\urcorner))]$
- (QC) $\forall q\forall\varphi\forall\psi[(\mathit{QuantSeq}(q) \wedge \mathit{Sent}(q\varphi) \wedge \mathit{SkForm}(q\varphi,\xi)) \rightarrow (\mathrm{Tr}(q\varphi) \leftrightarrow \forall\bar{a}\forall\bar{b}\mathrm{Tr}(\xi(\dot{a},\dot{f}(\bar{b}))))]$

Proof. (DC) Suppose $(M, Tr) \vDash Tr(\ulcorner\bigvee_{\varphi\in c}\varphi\urcorner)$. By EES we know that this holds if and only if $M \vDash^{*} \bigvee_{\varphi\in c}\varphi$. We thus have by the definition of $\vDash^{*}$ that this is the case if and only if $M \vDash^{*} \varphi_i$ for some φ_i in c. We conclude, again using EES, that this holds if and only if $(M, Tr) \vDash Tr(\ulcorner\varphi_i\urcorner)$ which is if and only if $(M, Tr) \vDash \exists\ulcorner\varphi\urcorner \in cTr(\ulcorner\varphi_i\urcorner)$.

(CC) Suppose $(M, Tr) \vDash Tr(\ulcorner\bigwedge_{\varphi\in c}\varphi\urcorner)$. By EES we know that this holds if and only if $M \vDash^{*} \bigwedge_{\varphi\in c}\varphi$. We thus have by the definition of $\vDash^{*}$ now that this is the case if and only if $M \vDash^{*} \varphi_i$ for all φ_i in c. We conclude, again using EES, that this holds if and only if $(M, Tr) \vDash \forall\varphi_i \in cTr(\ulcorner\varphi_i\urcorner)$.

(QC) Suppose $(M, Tr) \vDash Tr(q\varphi)$. By EES we know that this holds if and only if $M \vDash^{*} q\varphi$. Following the definition of $\vDash^{*}$ we have that this is the case if and only if $M \vDash^{*} \xi(\bar{x}, f(\bar{y}))$ for all substitutions. By EES this holds if and only if $(M, Tr) \vDash Tr(\xi(\dot{a},\dot{f}(\bar{b}))$ for all $\bar{a},\bar{b} \in M$. Therefore this holds if and only if $(M, Tr) \vDash \forall\bar{a}\forall\bar{b}\,Tr(\xi(\dot{a},\dot{f}(\bar{b}))$. □

We therefore know that the theory $\mathrm{CT}^-\!+ \vDash^*$ implies $\mathrm{CT}^- + \mathrm{I}\Delta_0$, the theory of CT^- with induction axioms for Δ_0 formulas from $\mathscr{L}_{Tr}$ due to a recent result by Enayat and Pakhomov (2018, Thm. 1). They show that that $\mathrm{CT}^-\!+$ DC is equivalent to $\mathrm{CT}^- + \mathrm{I}\Delta_0$, and since $\mathrm{CT}^-\!+ \vDash^*$ implies $\mathrm{CT}^-\!+$ DC, we know that $\mathrm{CT}^-\!+ \vDash^*$ also implies $\mathrm{CT}^- + \mathrm{I}\Delta_0$.

Theorem 2.5.1.4. $\mathrm{CT}^-\!+ \vDash^*$ *implies each induction axiom for* Δ_0 *formulas from* $\mathscr{L}_{\mathrm{Tr}}$*, the theory of* $\mathrm{CT}^- + \mathrm{I}\Delta_0$*.*

Enayat and Pakhomov show that $\mathrm{CT}^- + \mathrm{I}\Delta_0$ is equivalent to CT^-+DC and it is a new interesting open question whether a similar result holds for QC, i.e. whether $\mathrm{CT}^-\!+$ QC is equivalent to $\mathrm{CT}^- + \mathrm{I}\Delta_0$.

Question 2.5.1.5. *Is* $\mathrm{CT}^-\!+QC$ *equivalent to* $\mathrm{CT}^- + \mathrm{I}\Delta_0$*?*

Theorem 2.5.1.4 leads to the following corollary due to results by Cieśliński (2010a,b), Kotlarski (1986) and Enayat and Pakhomov (2018).

Corollary 2.5.1.6. $\mathrm{CT}^-\!+ \vDash^*$ *implies each of the following theories:*

- $\mathrm{CT}^- + \mathrm{DC}$
- $\mathrm{CT}^- + \mathrm{CC}$
- $\mathrm{CT}^- + \mathrm{I}\Delta_0$
- $\mathrm{CT}^- + \forall\varphi[\mathit{Prov}_{\mathrm{FOL}}(\varphi) \rightarrow \mathrm{Tr}(\ulcorner\varphi\urcorner)]$
- $\mathrm{CT}^- + \forall\varphi[\mathit{Prov}^{\mathrm{Tr}}_{\mathrm{Prop}}(\varphi) \rightarrow \mathrm{Tr}(\ulcorner\varphi\urcorner)]$
- $\mathrm{CT}^- + \forall\varphi[\mathit{Prov}^{\mathrm{Tr}}(\varphi) \rightarrow \mathrm{Tr}(\ulcorner\varphi\urcorner)]$
- $\mathrm{CT}^- + \forall\varphi[\mathit{Prov}^{\mathrm{Tr}}_{\mathrm{PA}}(\varphi) \rightarrow \mathrm{Tr}(\ulcorner\varphi\urcorner)]$
- $\mathrm{CT}^- + \mathit{Con}_{\mathrm{Tr}}$

Here we have that *Prov* is a formalised provability predicate, where the subscript FOL denotes 'in first order logic', Prop 'in propositional logic' and PA 'in Peano Arithmetic'. The superscript *Tr* denotes 'from premises contained in the satisfaction class'. Con_{Tr} denotes the formalised statement that the set of sentences contained in the satisfaction class is consistent. It is worth highlighting that $\forall\varphi$ ranges over all Gödel codes of sentences.

This conception of closing a satisfaction class under the extended T-schema is thus a natural semantic motivation for the strength and addition of these theories. Adding the extended T-schema to the compositional axioms for truth ensures that a multitude of nice properties for the theory follow.

This theorem has the following corollary that the extended T-schema is a non-conservative addition to the compositional axioms for truth. This follows from a result due to Wcisło and Łełyk (2017), fixing a gap in the proof offered by Kotlarski (1986) of this, that shows $\mathrm{CT}^- + \mathrm{I}\Delta_0$ is not conservative over PA since it proves the formalised consistency of PA. This is in contrast to the theory of the compositional axioms for truth and the theory of compositional axioms for truth with the regular T-schema.

Corollary 2.5.1.7. *The theory* $\mathrm{CT}^- + \vDash^*$ *is not conservative over* PA *or the* $\mathscr{L}_A$*-consequences of* CT^-*. In particular,* $\mathrm{CT}^- + \vDash^* \vdash \mathit{Con}(\mathrm{PA})$*.*

We therefore have that the extended T-schema is a strong condition to add to a weak theory of truth. The philosophical consequences of this can be seen in the following section.

2.6 Conclusion

This formal journey leads to the following conclusions. If one accepts that a theory of truth should apply to nonstandard syntaxes, for reasons I have argued in Section 2.2.2 or otherwise, and that the T-schema should hold for as many sentences not containing the truth predicate as possible, then the standard formulation of the T-schema is not sufficient. It does not capture enough of these sentences. It can be augmented with an extended T-schema by the use of a theory of semantics for nonstandard sentences, such as the Robinson inspired semantics in Section 2.4. This extended T-schema follows in the spirit of Tarski. It is a principle formed in the metalanguage, in terms of syntax, and one which inspires axiomatic principles such as those in Theorem 2.5.1.3. These principles, together with the basic compositional properties of truth, prove a variety of desirable properties over Peano Arithmetic and are non-conservative.

I propose that ensuring the extended T-schema holds, together with basic compositional properties of truth, should be taken as a new standard of minimal adequacy for a theory of truth. If a theory of truth cannot prove the extended

T-schema, then it is not able to capture the basic usage of the truth predicate. This is desirable on proof-theoretic grounds since it ensures conditions such as 'all theorems of the base theory (PA) are true', induction for simple (Δ_0) formulas involving the truth predicate, and the consistency of the set of true sentences.

The extended T-schema is an alethic motivation for these properties, however. Ensuring that the extended T-schema holds does not rely upon any further reflection about truth or its base theory, other than the disquotational nature of the truth predicate. Whereas one may argue that the proof-theoretic principles above requires reflection not about truth (such as belief in the consistency of PA and the extensibility of induction, for example), the extended T-schema does not make explicit reference to such commitments. Instead, it simply formulates the T-schema over arbitrary theories of syntax. This is a basic property of truth, together with a principle of neutrality towards what counts as sentences.

This has ramifications for the deflationist who takes conservativity as a commitment of their view, as Horsten (1995), Shapiro (1998) and Ketland (1999) have argued for. The nonconservativity of this theory of truth arises solely from considerations of the T-schema and truth's compositional properties. The T-schema seems to be a commitment of any theory of truth, let alone the deflationist for whom it is central, as Armour-Garb (2012) argues. Similarly, that truth commutes with the logical connectives and quantifiers seems to be a basic property of the truth predicate. These two truth-theoretic principles lead to nonconservativity, as stated in Corollary 2.5.1.7. This avoids counterarguments, such as Field's (1999), that previous nonconservative phenomenon rely on arithmetical consideration, as well as truth-theoretic consideration.

I therefore conclude that the deflationist about truth has a choice to make. They can either argue, *contra* Section 2.2.2, that we should only consider truth predicates over a fixed standard theory of syntax, or that conservativity of the truth predicate is not a commitment of their view.

For the deflationist who does deny conservativity, this result has some good news. One basic tenet of most varieties of deflationism about truth is that the T-schema allows one to derive all facts about truth. This shows that the T-schema actually has more power than one would first expect, when formulated as the extended T-schema. From this (together with the compositional axioms), one is able to derive the consistency of the set of true sentences and the ability to express everything provable in the base theory (PA) is true. This is formal support for the

deductive power the T-schema can offer. This offers a means for the deflationist about truth to argue for the adequacy of the eqvuivalence schema in accounting for all of truth's usages.

This new condition of minimal adequacy therefore implies negotiation of a trade-off for the deflationist. They can either accept the conservativity of truth, but must then also accept the strong assumption of a standard theory of syntax, or can reject conservativity and receive deductive power from the T-schema.

This choice for deflationists will motivate the next two chapters of this thesis. In order to decide which option is more suitable, we need to be clearer on what deflationism about truth means. Deflationism is presented as the thesis that truth has no 'substantive' nature, but what this means exactly, philosophically, is not entirely clear. This will be the subject of Chapter 3. Once we have an answer to this question, we will explore what what it means for a formal theory of truth to be deflationary, and this will be the subject of Chapter 4. This chapter shall answer the question above and argue that it does not entail conservativity and so deflationists can reject this and choose an extended T-Schema.

Open Questions

2.5.1.2 Is there a natural theory of truth $\mathrm{CT}^- + \mathrm{X}$, for some X, which is equivalent to $\mathrm{CT}^- + \vDash^*$?

2.5.1.5 Is CT^-+QC equivalent to $\mathrm{CT}^- + \mathrm{I}\Delta_0$?

2.4.2 Is there a natural way to extend Definition 2.4.1 to account for sentences containing a repeated nonstandard disjunction or conjunction formed with 'alternative' bracketing?

Chapter 3

Deflation beyond Disquotation: What is a Deflationary Theory of Truth?

This chapter is inspired by the choice left at the end of Chapter 2. Should the deflationist accept conservativity, but argue against applying the T-Schema to nonstandard models of syntax, or should the deflationist reject conservativity and boast the deductive power of a T-Schema? In order to answer this question, we need to be clear on what deflationism is and what it means philosophically for a theory of truth to be deflationary. It turns out that this question is not so easy to address, and there are a number of terms to precisify to understand the commitments of deflationism. This chapter will answer this question, so that the choice of Chapter 2 can be addressed in Chapter 4, where it will be argued that deflationists should reject conservativity.

Chapter Abstract

Deflationary theories of truth appear radically different alternatives to the traditional theories of truth, but it is not clear what sets them apart so distinctly. In this chapter I investigate how we should best understand deflationism about truth. I will conclude that we should understand alethic deflationism as the claim that a logical-linguistic-semantic theory of the word 'true' exhausts our understanding of truth, and that a deflationary property of truth is a pleonastic property. In order to establish this claim I argue against the understanding that all there is to a deflationary theory of

truth is some form of T-Schema, and against common refinements of what it means for a property of truth to be insubstantial. I conclude that, if I am correct, arguments against deflationism in the literature are in actuality only arguments against specific deflationary theories of truth, and deflationism about truth as a whole is able to withstand the force of these criticisms.

3.1 Introduction

Deflationary theories of truth appear radically different to the alternative theories of truth. Inflationary theories of truth focus on exploring what makes certain propositions true, the role of truth in philosophy, and the importance of truth in other disciplines such as science, mathematics and logic. The deflationist opposes that there is any deep exploration to be had here, and contests whether truth has any important part to play in our theories. Instead, she focuses on the linguistic role of truth and denies that there is any substantive nature to truth. Deflationists about truth often claim that truth has no underlying metaphysical nature,[1] has primarily linguistic roles[2] and should be regarded as something like a logical property.[3] For deflationists there is no question as to the nature of truth, for it is metaphysically thin.[4] The deflationists and inflationists disagree fundamentally over the metaphysical status of truth and its philosophical importance.

As common as they are, none of these statements I have just provided are particularly precise. As Wyatt (2016) points out, it is not even clear whether some of these claims are about the concept of truth, the property of truth or the word 'truth' itself. There are a great variety of 'inflationary' theories of truth,[5] which jointly oppose these claims, and offer a substantive analysis of truth. There is a disconnect between the two sides, but what exactly they disagree on has not been made explicit. What does the deflationist mean when she says that truth is insubstantial in opposition to the inflationist? It is this question that I aim to

[1]Horwich (1998) writes of truth that: "No wonder that its 'underlying nature' has so stubbornly resisted philosophical elaboration; for there is simply no such thing."

[2]Brandom (2002) provides "a sketch of the expressive role that is characteristic of the expression '...is true'."

[3]Field (1994c) argues that: "the word 'true' has an important logical role."

[4]Shapiro (1998), for instance, characterises deflationism as "metaphysically thin, or natureless, or lightweight".

[5]For example: correspondence theories, identity theories, pragmatist theories, coherence theories, and pluralist theories.

address here, by providing a clear understanding of what it is about these theories of truth that makes them deflationary. I want to be specific about what it is that deflationists, as a whole, are endorsing when they argue for deflating the philosophical study of truth.

It should be acknowledged that deflationism is a term of art, and my purpose is to regiment its use as much as to classify it. The term could well be flawed and our intuitions about its use may not all be jointly satisfiable. With that being acknowledged, my methodology in exploring this question will be driven by existing examples of deflationary theories of truth. We have a number of theories which are agreed to be, and proposed by their authors as, deflationary theories of truth, and similarly a number of inflationary theories of truth widely agreed not to be deflationary. These will be my tests for any criterion of alethic deflationism – if a criterion states that an inflationary theory is deflationary, or vice-versa, then that will bring the criterion into question.

My understanding of alethic deflationism here is wide and includes all theories of truth which deny metaphysical substance and powerful explanatory roles to truth, those whose authors often describe themselves as deflationary. A common criterion of deflationism is that the theory states that all there is to truth is captured by some form of T-Schema. This is a scheme of formulas of the form "'S" is true if and only if S' where "S" is an expression of S. Whilst this may be one terminological usage of 'deflationary', in Section 3.2 I shall argue that deflationism about truth should be conceived as far wider than this narrow classification. Deflationists can endorse theories where a T-Schema is not fundamental to their theory.[6] In Section 3.3 I shall explore alternative proposed criteria of what it means for a truth property to be insubstantial, and argue that these are similarly inadequate. In Section 3.4 I shall argue positively for the claim that we should understand a deflationary theory of truth as a logical-linguistic-semantic theory of the word 'true', and that deflationism is the thesis that such a theory exhausts our concept of truth. This results in the understanding that a deflationary truth property is a pleonastic property, in the sense of Schiffer (2003), and this is a good clarification of the claim that a deflationary truth property is insubstantial. I will conclude that this understanding of alethic deflationism defends deflationism as a

[6]In Chapter 2 I viewed the T-Schema as a necessary condition for any theory of truth. This position will not be challenged within this chapter, and I will instead argue that it is not the sole condition for a theory of truth to count as deflationary.

whole from opposing arguments in the literature, since these are ultimately only arguments against a T-Schema.

3.2 Beyond Disquotation

Semantic ascent and descent appears to be one of the primary linguistic functions of the word 'true'. Semantic ascent and descent refers to the capacity of a truth predicate to affirm semantic content from an 'object language' within a 'meta language' (ascent) and vice-versa (descent). An object language is any language consisting of terms referring to objects, properties of the objects, and relations between them. The meta language is the language which discusses sentences belonging to the object language, where such sentences are the objects of the meta-language. One of the predicates of this meta language is the truth predicate, which is a vehicle between the two languages. For example, snow is an object and 'is white' is a property of snow. We can denote that snow has the property of being white in the object language by stating "snow is white". We can instead endorse this in the meta language by stating that the sentence 'snow is white' is true. Similarly, given a sentence "'S" is true', where 'S' is an object in the metalanguage, we can endorse S in the object language. Semantic ascent and descent is often stated in one of two ways: as an equivalence schema or as a disquotational schema.

The equivalence schema (ES) is all instances of the form:

$$\langle \text{P} \rangle \text{ is true if and only if P}$$

where $\langle \text{P} \rangle$ is a schematic variable which ranges over all propositions and an instance $\langle \text{P} \rangle$ is the proposition that P. The disquotational schema (DS) is all instances of the form:

'S' is true if and only if S

where 'S' is a schematic variable which ranges over all sentences and an instance 'S' expresses S.

The difference in name between these two schemes is their notion of truthbearers: propositions for ES and sentences for DS. I shall refer to these as the T-Schema (TS), since my remarks do not depend upon specifying either interpretation of truthbearers, although I will refer to sentences as the bearers of truth

for ease. Other notions of truthbearers, not just propositions but also beliefs, utterances, etc. should be able to be substituted in here freely.

As Armour-Garb and Beall (2005, p. 2) note, a T-Schema is widely accepted amongst philosophers.

> At least since Tarski (if not since Aristotle), most philosophers have taken the following equivalence schema to be central to our concept of truth.

Many philosophers go further, however, and a T-Schema is not just central to their concept of truth, but forms their theory of truth entirely. For example, Horwich's minimal theory of truth consists of all non-paradoxical instances of ES (Horwich, 1998). Field (1994c) defends a disquotational theory of truth, which rests upon the disquotational schema. Beall (2009) defends a transparent theory of truth, where 'α' is true and α are inter-substitutable in all non-opaque contexts. Künne's modest theory of truth is the universal propositional quantification of all instances of ES (Künne, 2003).

These philosophers have more than a reliance upon a T-Schema in common, they also all hold deflationary views of truth. In fact, many deflationists about truth maintain a T-Schema, or its instances, as the only theory of truth that we require and argue that a T-Schema explains everything we need to say about truth. With many of the high-profile deflationary theories of truth exemplary in this regard, it is often understood that the T-Schema is not just fundamental to many deflationary theories of truth, but entirely characterises what deflationism is. Armour-Garb (2012, p. 2), for example, defines T-deflationists (truth deflationists) in just such a manner:

> what distinguishes T-deflationists from T-inflationists is that only the former take instances of (TS) to be *fundamental*, both conceptually and explanatorily ... The instances of (TS) are conceptually fundamental in that they do not follow from definitional relations holding among the concept of truth and more basic concepts in terms of which 'true' can be defined ... the instances of (TS) are *fundamental explainers* of truth-talk in that everything that we do with the truth predicate can be explained, ultimately in terms of the instances of (TS).

Soames (1998, p. 231) follows a very similar line when characterising deflationism:

> This brings us to the leading idea behind deflationism about truth - namely that claims of the sort ⌜It is true that S⌝ and ⌜The proposition that S is true⌝ are trivially equivalent to S and that this equivalence is in some sense definitional of the notion of truth.

Eklund is in agreement with this conception. He writes that the exhaustion thesis is the "basic claim" (Eklund, 2017, p. 3) of deflationism, where this is the claim that "What truth is, is exhausted by some schema like (ES) ... or (DS)" (Eklund, 2017, p. 2). Similarly Field (1994b, p. 405) writes that: "'Deflationism' is the view that truth is at bottom disquotational." Lastly, Armour-Garb and Beall (2005, p. 3) agree in their chapter *Deflationism: The Basics*:

> What distinguishes deflationists from substantivists – what constitutes the heart of deflationism – is that deflationists take the instances of (ES) to be *fundamental*, both conceptually and explanatorily ... The instances of (ES) are bedrock.

It appears to be a commonplace understanding that deflationism is the thesis that all there is to truth is some form of a T-Schema, and that this is all there is to a deflationary theory of truth. In this section I wish to push back against this conception of deflationism, and show that deflationism about truth can, and should, be understood as wider than a T-Schema. Alethic deflationists can happily admit aspects of truth that extend beyond a T-Schema, and cannot be explained by a T-Schema, but retain their insubstantial stance on truth. Further, deflationists can even build their theory of truth on other grounds, and (if they so desire) derive the T-Schema from this: a T-Schema does not need to be the foundational bedrock of a deflationary theory of truth.

One main piece of evidence supporting this position is that (at least) two deflationary theories of truth do not place a T-Schema in central position. Grover, Camp, and Belnap's (1975) prosentential theory of truth, more recently advocated by Brandom (1994), is a theory of truth in which the word 'true' is not a property-ascribing predicate, but a prosentence-forming operator. Here, truth is understood purely as an expressive linguistic resource. Typically the word 'true' is viewed as a 'property-ascribing' predicate, which ascribes a property (truth) to truthbearers (such as sentences). Prosententialists deny this, and instead, they argue that the word 'true' behaves as an operator, which produces prosentences.

In English, pronouns such as 'she', 'he' and 'it' can stand in for objects, such as in the sentence "Jane wanted to go out with her friends, but she was too busy" where 'she' stands in for 'Jane'. Prosententialists argue that the truth predicate provides a similar linguistic role and 'it is true' forms a prosentence, which can stand in for a previously mentioned sentence. For example, in the sentence "Jane's friends do not believe that she was busy, but it is true" the 'it is true' stands in for the sentence 'Jane was busy.' Prosententialists deny that 'is true' is really a predicate, and instead treat it as an operator, much like how modal logic treats 'is possible' as an operator rather than a predicate. Advocates of prosententialism describe their theory as deflationary, since they deny any substantive nature to the property of truth and, quoting Brandom (2002, p. 117), "preclude one from treating the notion of truth, and hence of truth conditions, as explanatory raw materials". Yet, prosententialists do not treat a T-Schema as central, but instead can derive its instances from their prosentential account of truth: '"S" is true' is a prosentence that can stand in for the sentence S, and hence they are logically equivalent in the correct context. The prosentential account of truth treats 'S is true' differently to a T-Schema and adds an anaphoric link to the expression 'S is true' that goes beyond the T-Schema. The T-Schema is certainly not fundamental to prosententialists, either conceptually or fundamentally, and is instead explainable from their wider account of truth as a prosentence-forming operator.

Strawson's (1948) performative theory of truth also does not give the T-Schema a fundamental role and even denies it as a feature of truth entirely. Strawson objects to those who treat truth as a meta-linguistic predicate of sentences, one which ascribes a semantic property of truth to a sentence in an object language, and instead highlights its assertive or performative role. In particular, Strawson highlights that the usage of 'that's true' requires a linguistic occasion to have already taken place to be meaningful, for otherwise it would be nonsensical. Analogous to someone uttering 'ditto' as the first statement of a conversation, an utterance of 'that's true' without any previous statements would make no sense. Strawson concludes that 'that's true' should not be viewed as equivalent to making the statement it refers to, nor as about what has been said previously. Instead, it is a performative linguistic act in which the speaker agrees with a statement. Truth is not a description of a sentence, but an endorsement of one. Strawson denies the T-Schema, and even the semantic ascent/descent analysis, as a feature of truth, since it is mistaken in both treating 'is true' as a statement about a statement and

'S' is true as equivalent to the statement S. Strawson's analysis is deflationary, however, since he denies a substantive property of truth and denies truth plays an important explanatory role in our language beyond endorsing previous statements. Strawson provides a linguistic theory of truth in which the word 'true' is given no deeper meaning beyond its performative role.

These theories highlight that deflationism can be understood as wider than a T-Schema, and that not all deflationists believe that it is fundamental to their theory of truth. There are further reasons that one might think deflationism should be conceived of beyond a T-Schema, however, even for those who give it a highly important place within their theory of truth. For example, some deflationists may argue that truth has linguistic features which go beyond the T-Schema: in particular its compositional nature. In (most) cases, given two sentences 'S_1' and 'S_2' and a connective between sentences R we have that:

'S_1 R S_2' is true if and only if 'S_1' is true R 'S_2' is true

This is near uncontroversial when R is a logical connective such as '$\wedge$' or '$\vee$', but also seems true of other connectives such as 'because', 'while', and 'necessitates'. This compositional feature of truth can be seen as part of its linguistic function, and is certainly not one denied by prominent (T-Schema fundamental) deflationists, such as Horwich (1998), Field (2008) and Künne (2003). Yet, Tarski (1956, p. 257) shows, in a formal framework, that general rules of the form:

For all sentences 'a' and 'b': if 'a' is true and 'b' is true, then '$a \wedge b$' is true

cannot be derived from the T-Schema alone. In fact, Tarski (1956, §5, Thm. 2) showed that taking compositional rules in combination with a method of determining truth for atomic sentences can imply the T-Schema instead. In the face of this result, Quine (1986, p. 41) (a deflationist himself) takes compositional axioms as a fundamental part of his theory of truth. A Tarski-Quine-style theory of truth, where truth is understood by a combination of compositional rules and an 'insubstantial' method (T-Schema or otherwise) for determining the truth of atomic sentences does not treat the T-Schema as either conceptually or explanatorily fundamental, but is still deflationary in nature. The T-Schema plays an important part of such a theory, but no more important than general compositional rules for truth which are not reduced to the T-Schema.

Another reason that a deflationist might consider the truth predicate as understood beyond the T-Schema comes from consideration of 'general sentences'. A general sentence, such as 'everything provable in arithmetic is true' cannot be directly analysed using a T-Schema, since this results in the ungrammatical fragment 'everything provable in arithmetic'. The standard approach to general sentences is to understand them as an infinitely long conjunction which is impossible to state in natural language. Deflationists, such as Quine (1986), Horwich (1998), and Field (2008), argue that this becomes one of the main expressive functions of the truth predicate. For example, 'everything provable in arithmetic is true' is understood as equivalent to:

> 'P_1' is provable in arithmetic and 'P_1' is true and 'P_2' is provable in arithmetic and 'P_2' is true and 'P_3' is provable in arithmetic and 'P_3' is true and ...

where 'P_1', 'P_2', 'P_3', ... enumerate the true arithmetical propositions. This infinitely long sentence, following the T-Schema, is then equivalent with the infinitely long 'True-free' sentence:

> 'P_1' is provable in arithmetic and P_1 and 'P_2' is provable in arithmetic and P_2 and 'P_3' is provable in arithmetic and P_3 and ...

It is claimed that since we cannot state this infinitely long sentence, we introduce the truth predicate to express it as 'everything provable in arithmetic is true'. Yet, for a deflationist who objects to the usage of infinitely long sentences, general sentences are a problem for a T-Schema. If infinitely long sentences are not admissible here, then the deflationist cannot understand general sentences in this way, and potentially needs resources beyond a T-Schema to discuss these.

Azzouni (2001) is one example of a deflationist who takes this stance, and he views the T-Schema as derived from (but not fundamental to) his theory of truth. Azzouni criticises the standard deflationary manoeuvre of translating a general sentence as an infinitely long conjunction, and argues that it requires a grasp of an infinite set of expressions, which we cannot do.[7] In the above example, the infinitely long sentence contains each statement provable in arithmetic, a collection that Azzouni argues cannot be grasped. Azzouni instead develops a theory

[7] Azzouni notes that we can, and do, grasp things about infinite collections, but not the whole collection together.

of truth which rests upon an 'anaphorically unrestricted' quantifier – a quantifier whose variables can appear in either sentential (quoted) or nominal (named) contexts. This allows one to state 'everything provable in arithmetic is true' not as an infinitely long sentence, but as:

For all p: If 'p' is provable in arithmetic, then p

Where p ranges over all sentences. This anaphorically unrestricted quantifier captures the content of 'everything provable in arithmetic is true' without reliance upon a T-Schema or grasp of an infinitely long set of sentences. Azzouni shows that the theory is then able to derive each instance of the T-Schema, but he does not give the T-Schema priority and instead treats the quantifier as fundamental. Azzouni describes himself as a deflationist about truth and endorses "that 'true' plays a humble expressive role that facilitates communication, and that's *all* it does" (Azzouni, 2006, p. 10). Azzouni is therefore a deflationist who denies that the T-Schema is a foundation for his theory of truth.

One final concern with taking the T-Schema as constitutive of a deflationary theory of truth is that typically classical theorists of truth do not admit all instances of the T-Schema, and deciding which instances to accept appears to be a feature of truth which goes beyond the T-Schema. Classical deflationists often restrict the T-Schema due to consideration of the semantic paradoxes, such as the Liar sentence L which expresses "'L' is false". If we substitute this into the relevant instance of the T-Schema we derive that 'L' is true if and only if 'L' is false (and hence L if and only if not-L) – a contradiction. There are two main deflationary responses here. One option is to move to a non-classical logic, where 'L' has a non-classical truth value, and still admit all the instances of the T-Schema; Beall (2009) takes this route, for example. An alternative response is to keep classical logic, but restrict the T-Schema in some way. Horwich (1998) takes this route, and specifies that we only admit a maximal consistent set of T-Sentences (instances of the T-Schema). McGee (1992, Thm. 2) proves in a formal framework, however, that there are infinitely many incompatible maximally consistent sets of T-Sentences, and worse, that none of these are recursively axiomatisable (informally, there is no in-principle computable procedure which can determine whether any T-Sentence is a member of such a set or not). McGee concludes that solely requiring a theory of truth to generate a maximal consistent of T-Sentences

is too weak, since we cannot hope to actually construct such a theory.[8] It appears that building a classical theory from T-Sentences alone requires some additional resource to choose exactly which T-Sentences to pick, and such a resource will have to go beyond the T-Schema (whether there exists such a resource which is deflationary in nature I leave as an open question). This resource will be an aspect of truth that cannot be reduced to the T-Schema.

Such examples show, I hope, that we should not view the T-Schema as all there is to a deflationary conception of truth. A deflationist can (and perhaps should) admit uses of the truth predicate which go beyond the T-Schema, and can, if they wish, build a theory of truth where the T-Schema is derived, rather than fundamental. The question then remains of just what a deflationary conception of truth is, if it is more than, or different to, the T-Schema? In the next section I shall tackle conceptions of deflationism which focus on understanding the metaphysical claims of deflationism, and argue that whilst these elucidate what it means for truth to be 'insubstantial', they do not characterise deflationism about truth either.

3.3 An Insubstantial Truth Property

In the previous section I argued that deflationism about truth should be understood as wider than the thesis that all there is to truth is some form of T-Schema. In this section I would like to interrogate what it means to be a deflationist about truth more deeply. For instance Brandom (2002), Strawson (1948) and Azzouni (2006) all consider themselves deflationists about truth, but do not take the T-Schema as central to their theory of truth. Why is it that their theories of truth are considered deflationary, then?

One common way to answer this question is by examining the metaphysical conception of a deflationary truth property. Often deflationists claim that the truth property is not a deep and significant metaphysical property, but instead "insubstantial" (Horwich, 1998, p. 52), "thin" (Armour-Garb and Beall, 2005, p. 1) or "lightweight" (Shapiro, 1998, p. 495) with no "underlying nature" (Horwich, 1998, p. 2). This is in comparison to inflationary theories of truth, where the property of truth is a genuinely deep and metaphysically robust property. Whilst such adjectives may suggest what the deflationist has in mind, they are certainly

[8] McGee also concludes that this requirement is too strong, since even if we could construct such a theory, it would not capture all the useful features of our ordinary notion of truth.

not precise rigorous notions, and if we are to understand what deflationism is these need to be brought to light.

I should first provide the caveat that I do not consider such claims to be fundamental to deflationism. As I have stated, there are deflationists about truth (prosententialists) who deny that the word 'true' is a property-ascribing predicate. For such deflationists, remarks on the insubstantiality of the truth property do not characterise their views, for they deny that there is such a property. There are, however, a great many deflationists who endorse a property of truth. I believe that considerations of what they mean by this are instrumental in understanding deflationism more widely. I will explain this point in the following section, where I shall focus on deflationism about truth as fundamentally a logical-linguistic-semantic theory about the word 'true'. My aim in this section is to consider current approaches to understanding what it means for a truth property to be insubstantial, and argue that, as they stand, they are inadequate even for the property-endorsing deflationists. I will consider three such claims: Damnjanovic's (2010) argument that a deflationary truth property is a revelatory property; Edwards' (2013a) claim that a deflationary truth property is an abundant property and Wyatt's (2016) claim that a deflationary truth property is one lacking a constitution theory. My test for this will be common examples of deflationary and inflationary theories of truth, and I will show that each claim mistakenly classifies one or more of these.

Damnjanovic (2010) argues that a deflationary truth property is one which is both a logical property and a revelatory property (also referred to as a transparent property, within the literature). What, exactly, a logical property is requires far more attention than I provide here to do justice,[9] but some common examples of logical properties are conjunction, quantification, and equality. On the other hand, common examples of nonlogical properties are charge, addition and colour. Given these examples, it appears that no deflationary truth can be a purely logical property, due to it being a predicate of truthbearers in the metalanguage, and not simply objects in either language. This is in contrast to logical properties like quantification and equality which are predicated of all objects and not specifically truthbearers. This is something deflationists have admitted, for example, Künne (2003, p. 338) writes that: "it appears reasonable to call truth a broadly logical

[9]Formal approaches to defining logicality are discussed in Chapter 4, Section 4.4, where it is argued that under at least some natural understandings of logicality formal truth properties are not logical properties.

property. (Only '*broadly* logical, because the concept of a proposition is not a logical concept)". Further, Horsten (2011, p. 65) argues that even in formal theories of truth the truth predicate should not be regarded as solely logical, but logico-linguistic, since the bearers of truth are meaningful sentences.

I will hence instead focus on Damnjanovic's argument that a deflationary truth property is a revelatory property. This is understood to mean that the concept of truth *reveals* the property of truth, which means that by grasping the concept of truth one is in a position to grasp the nature of the property of truth, without further empirical or *a priori* reasoning. An example of this is conjunction: grasping the concept of conjunction reveals the property of conjunction immediately, without any other investigation into the nature of conjunction required. In contrast, one can grasp the concept of water, but still not be in position to grasp that the property of water is H_2O without further empirical research. For Damnjanovic, the deflationary claim is that the concept of truth reveals the property of truth. Once one understands the concept of truth, then the property of truth is revealed, without any further investigation or metaphysical research required.

I believe that many deflationists would probably be sympathetic with this claim, but that this should not be regarded as the foundation of their view. Eklund (2017, §3) takes aim at such a characterisation. In a remark Eklund ascribes to Raatikainen, he argues that if the correspondence theorist holds that our concept of truth is a correspondence concept, and this is the nature of the truth property, then a correspondence property of truth is a revelatory property. Similarly, we can see that if the coherence theorist of truth holds that our concept of truth is a coherence concept, and this is the nature of the truth property, then a coherence property of truth is a revelatory property. This criterion depends upon the concept of truth held and does not distinguish between properties, but merely different epistemic concepts of the properties. As Edwards (2013a, p. 283) observes, even if this criterion provided the correct answers, the revelatory hypothesis is an epistemic, rather than metaphysical claim, and is a claim about our understanding of the property of truth rather than the actual nature of the property of truth itself. The criterion does not tell us what is insubstantial about the property of truth, but instead about the concept of truth being discussed.

Edwards (2013a), having rejected the revelatory hypothesis, argues for an alternative distinction between deflationary and inflationary properties of truth, utilising Lewis' (1986) distinction of sparse and abundant properties. Lewis grades

properties with a partial ordering, where properties at one end are considered sparse and properties at the other end abundant. The most sparse properties are those most natural and fundamental to the world, whereas those properties on the abundant end are gerrymandered and artificial. For example, physical properties of particles like charge and spin, which are fundamental to reality, fall on the extreme of the sparse end. We then have other relatively sparse properties, defined from the most sparse properties, such as molecules and other natural kinds. On the other end of the scale, we have abundant properties – artificial and unnatural properties such as Goodman's grue and bleen[10] or a lengthy enumeration of arbitrary objects. Edwards argues that a deflationist should accept that the truth property is an abundant property on this distinction, whereas the inflationist would claim that truth is more of a sparse property. The Lewisian framework enables one to formulate the deflationary claim that truth is insubstantial, since it is an artificial property which is neither metaphysically special nor fundamental to the world.

Wyatt (2016, §III.4) analyses the adequacy of the sparse/abundant distinction for truth properties. He notes that there are different criteria of abundance, but each of these has shortcomings. For example, one measure of sparsity is that P is sparser than Q if the length of chain of definability of P from the sparsest properties is shorter than the length of chain of definability of Q from the sparsest properties. Applying this precisely to different theories' truth properties is no easy task, but it appears that certain inflationary correspondence properties will come up as on the abundant end of the spectrum. Wyatt notes that the property of 'being isomorphic to a worldly fact' will certainly have a very long definability chain from the fundamental physical properties, making a correspondence truth property an abundant property. This is certainly not what we want our criterion to say about a canonical inflationary truth property.

Another criterion of sparsity that Wyatt notes is that P is sparse if it has causal-explanatory power. As Wyatt observes, that a deflationary truth property is abundant under this account merely follows from the special deflationary claim that truth lacks *any* explanatory power. This criterion of deflationism reduces to an already familiar epistemic distinction, which whilst certainly characteristic of deflationism, does not clarify the particular metaphysical claims of deflationism.

[10]Goodman (1955, p. 74) defines these as the property of being green until time t, and blue afterwards, and the property of being blue until time t, and green afterwards, respectively, where we understand t as a term for a specific time.

This distinction is intended to explain what is insubstantial about deflationary truth properties, not the different epistemic roles of deflationary and inflationary truth theories.

The final criterion of sparsity that Wyatt notes is that P is sparse if it grounds genuine similarities among its members. Wyatt notes that this follows from the deflationary claim that truth lacks in exlanatory power and hence is not definitional of insubstantiality, as claimed in the previous paragraph, but I think that there is a deeper worry here. Taking a minimalist truth property as an example, the minimalist who endorses grounding appears in a good place to endorse that truth does ground a similarity amongst its members. The similarity being they all express a sentence 'S' such that S. Truth is the grounds for this similarity, for the minimalist, but it would be wrong to conclude that their truth property is inflationary from this.

Wyatt (2016), having rejected the previous approaches, understands a deflationary truth property to be one which lacks a constitution theory, and specifically argues that a deflationary property of truth lacks any constitution theory which is not revealed by the concept of truth.

A constitution theory for a property P is a set of propositions C such that:

> For every x, x instantiating P consists in x instantiating C

Wyatt argues that the deflationary thesis is that there is no metaphysical essence of truth, C, where C is what truth is constituted in. Wyatt further specifies the restriction that C cannot be revealed by the concept of truth (in the sense of Damnjanovic earlier) since some deflationary theories of truth do admit revelatory constitution theories, but these should not be counted as genuine constitution theories.

My issue with this distinction is that an inflationary primitivist truth property is going to come out as deflationary on this approach – whether we add the revelatory restriction or not. A primitivist about truth endorses a substantive truth property which plays an important philosophical role, but also that no account of the property can be given and it must be treated as a primitive notion. Moore (1899) and Russell (1904) are taken to be original advocates of this position, but this has been more recently endorsed by Asay (2013). For the primitivist, no possible theory of the property can be given, and hence they admit no constitution

theory. The constitution theory criterion views at least one major substantive property of truth as deflationary.

These three hypotheses appear to be the main contenders for understanding what it means for a deflationary truth property to be insubstantial, but if I am correct, then none of these are satisfactory. Further, since each focuses upon a deflationary truth property, none of these are suitable for understanding deflationism at large, where we have theories which admit no truth property at all. In the next section I will argue that we should conceive of deflationism about truth as primarily the thesis that all one requires for a theory of truth is a logical-linguistic-semantic theory about the behaviour of the word 'true'. I shall then argue that if this theory does endorse a truth property, then we should understand that this is a *pleonastic* property. Our efforts here will not be wasted, however, for I aim to show that the three criteria on offer here can be derived from this, given an appropriate interpretation of 'abundance' and 'consists'.

3.4 Logical-Linguistic-Semantic Theory of 'True'

For my characterisation of deflationism about truth, I wish to move away from focussing on what it is to be a deflationary truth property. As I have previously stated, not all deflationists endorse a truth property, and even those that do endorse a property do not give it their focus. Instead, deflationists focus on the behaviour of the word 'true' and provide a theory of truth by giving a theory of the word 'true'. In this section I will argue that alethic deflationism is the thesis that all we need for a theory of truth is a theory of how the word 'true' behaves, and that for deflationists this behaviour is logical-linguistic-semantic in nature. I will demonstrate that many common deflationary theories of truth, as well as those uncommon theories I considered in Section 3.2, fall into this category. I will then argue that those theories which admit a deflationary truth property understand this to be a *pleonastic* property in the sense of Schiffer (2003), and that this provides an adequate articulation of the claim that the property of truth is insubstantial.

Traditional theories of truth typically provide an account of truth by understanding its metaphysical or epistemic features. For example, a correspondence theory of truth understands a truth property to pick out a relation between truth-bearers and some form of metaphysical facts and a coherence theory of truth

understands the truth property to connect a network of truthbearers. For either theory it is this truth property that is of interest, and the word 'true' has utility because it expresses this important property of truth. Other substantive theories of truth are similar, for example a pragmatist theory of truth identifies a truth property with pragmatic success, and the word 'true' expresses this. The deflationist, on the other hand, does not consider the truth property as the fundamental part of their theory of truth, from which their wider account of truth is derived. Deflationary theories of truth instead focus on understanding the function of the word 'true' and derive their metaphysics (if any) of the truth property from this behaviour. The deflationist provides an account of linguistic correctness for using the word 'true' and denies that there is any deeper significance to truth than their linguistic account.

A key example of this is theories of truth which rest upon a T-Schema: such as Horwich's (1998) minimalist theory of truth, a disquotational theory of truth or Künne's (2003) modest theory of truth. Taking Horwich's minimalist theory of truth as an example, this is a theory of truth consisting solely of all non-paradoxical instances of the equivalence schema: $\langle P \rangle$ is true if and only if P, where $\langle P \rangle$ denotes the proposition that P. Each of these instances provides an account of when it is correct to predicate the word 'true' of a particular proposition. Taken together, the schema as a whole provides a complete account of the rules of use of the word 'true'. We predicate it of propositions $\langle P \rangle$, and we predicate 'is true' of $\langle P \rangle$ if and only if P.

This linguistic account of 'true' as an equivalence schema is compatible with an inflationary theory of truth. For example, a correspondence theorist would argue that truth is the property of correspondence with facts. They endorse that the predicate 'is true' expresses this property and can derive from this each instance of the equivalence schema as an account of when it is correct to state a particular proposition as true. This does not make their account deflationary, however, since this is derived, rather than fundamental. The inflationary/deflationary distinction is a matter of priority and substance: the inflationist offers a theory of the truth property, and then derives information about the predicate from this. The deflationist offers a theory of the word 'true' first, and treats this as their theory of truth. For the deflationist, no deeper or more substantive theory of truth is required, and no special account of the property of truth is needed. For the deflationist, an account of the word 'true' *exhausts* their theory of truth and all

other features of truth can be explained in terms of this account. The inflationist offers a far wider theory of truth, that discusses things such as the metaphysical or epistemic role of the truth property.

It should be noted that it is not enough to classify deflationism as a theory of truth which provides an account of the word 'true'. Deflationary accounts of the word 'true' are *logical-linguistic-semantic* (LLS) in nature. They define the function of the word 'true' using solely logical, linguistic and semantic concepts – concepts such as equivalence, truthbearers, quantification, reference and anaphora. This is certainly a broad category, but narrow enough that it excludes metaphysical, epistemic and normative concepts, including typical inflationist notions which fall into these such as facts, correspondence, coherence and success.

I contend that a deflationary theory of truth is a logical-linguistic-semantic account of the word 'true' and that for the deflationist such an account derives all that we need say about truth. I stated at the beginning that my test of correctness for such a classification will be common examples of deflationary and inflationary theories of truth, and I hope to provide a number here that demonstrates my account passes this test. I sketched earlier that Horwich's minimalist theory of truth consists solely in instances of the equivalence schema - each instance providing a rule of use for the application of the truth predicate to a particular proposition. This is an LLS account of the word 'true' since it relies solely on the concepts of propositions, equivalence, and reference. For Horwich, this account exhausts all that needs saying about truth, and is used to derive all features of truth in need of explanation.

Künne's modest theory of truth is another example of a theory that fits into this criterion neatly. Künne's theory is neatly stated with the single axiom: $\forall x[x$ is true $\leftrightarrow \exists P(x = \langle P\rangle \wedge P)]$. The modest theory of truth is a quantification of all instances of the equivalence schema and again tells us when it is correct to prescribe 'is true' to an object. The axiom tells us that this is correct when the object expresses a proposition P and when P is the case. The modest theory is a theory of the word 'true' and uses concepts of reference, equivalence, propositions and also (propositional) quantification – all logical, linguistic or semantic notions. As with Horwich, for Künne we can derive all features of truth required from the modest theory.

Theories of truth based upon a T-Schema may exemplify my understanding of deflationism, but the alternative deflationary theories of truth I considered in Sec-

tion 3.2 fit this just as neatly. Prosententialism, for example, is similarly a theory about the behaviour of the word 'true'. As discussed in Section 3.2, prosententialism is a theory detailing the word 'true' as a prosentence-forming operator. It is a theory explaining the word's linguistic role using concepts from linguistics and logic such as prosentences, equivalence and substitution. The prosentential theory of truth denies that there is anything more to be said about truth, or equivalently, endorses that their logical-linguistic-semantic theory of the word 'true' exhausts truth. This is the deflationary aspect of their theory of truth, and falls neatly into my classification of deflationism.

Strawson's performative theory of truth is similarly a theory about the LLS behaviour of the word 'true' as well. Strawson provides a theory of the phrase 'that's true' and its behaviour in language, rather than any property or concept of truth. He analyses the phrase's conversational role using concepts from linguistics and logic and presents this as all that needs to be said about truth and its role in philosophy and other areas of inquiry. For Strawson, a theory about the linguistic function of the word 'true' exhausts our understanding of truth, and this coheres neatly with my conception of deflationism.

This understanding of deflationism also rules out substantive theories of truth for which it is sometimes tricky to analyse why they are inflationary. A correspondence theory of truth is inflationary, on my account, since it provides an analysis of the property of truth, rather than the word 'true'. Further, this analysis is not LLS in nature, but metaphysical, instead. Similarly, a pragmatist theory of truth comes out as inflationary on my account. This theory provides an analysis of the property of truth as well, and this analysis is epistemic in nature, rather than metaphysical.

These test cases may be easily identified, but we saw in Section 3.3 that a primitivist theory of truth is often hard to analyse as substantive. This theory denies that it is possible to provide a theory of truth, but endorses that the concept of truth has a useful explanatory role for philosophy as a primitive notion. This theory is not deflationary because it fails to provide an LLS theory of the word 'true' and inflationary because it states any provided theory of the word 'true' cannot exhaust our understanding of truth. Under my conception of deflationism, even awkward substantive theories are easily classified as inflationary.

One interesting test for the view is the coherentist theory of truth. One could provide a coherency theory of truth whereby a truthbearer is true if it is consistent

with a network of truthbearers, or logically entailed by them. Such a theory would provide an analysis of truth that is logical-linguistic-semantic and therefore, under my view, this theory would actually be deflationary in nature. I do not see this as a challenge to my understanding. Such a coherence theory of truth would appear to be deflationary, since its analysis of truth ascribes no deep nature beyond simple logical notions to truth. Crucially, such coherence theories of truth are not commonly advocated because of this. Logical entailment appears too weak to derive all the truths we endorse and there exist (more than) two sentences which are both consistent with collections of truthbearers, but inconsistent with each other. Coherence theorists hence endorse much stronger notions of coherence, such as explanatory entailment, which are not logical-linguistic-semantic in nature. These coherence theories of truth are not deflationary under my view as they instead make use of metaphysical or epistemological resources.

I thus contend that we should identify deflationary theories of truth as logical-linguistic-semantic theories about the behaviour of the word 'true'. These are not theories which introduce the word 'true' as expressing a particular property of truth, and then explore what this property is, but instead take the the word 'true' as fundamental to understanding truth and examine how it behaves linguistically. Deflationists focus on the word's LLS behaviour and the expressive utility it provides, rather than its ability to express a certain property and the utility of this property. Thus far this is not incompatible with a substantial property of truth, but deflationists add the additional claim that this is the only account or theory of truth that we require. Deflationism is the thesis that a theory of the logical-linguistic-semantic behaviour of the word 'true' *exhausts* our understanding of truth. Such a theory of the word 'true' can be used to derive anything further we need to say about truth, and there is no need for a theory of the property or concept of truth beyond this.

Many deflationists do admit, however, that there is a property of truth. This property does not have an important role in their theory, is not useful explanatorily and is claimed to be metaphysically insubstantial. The question remains, from Section 3.3, on what it means for this property to be insubstantial. I submit that a natural understanding of this, for deflationists who admit a property of truth, is that this is a *pleonastic* property. Schiffer (2003, §2.3) defines a pleonastic property as one which (only) results from a 'something-from-nothing' transformation – where we move from a sentence of the form 'a is P' to the sentence 'a has the

property of being P'. In other words, a pleonastic property is one which arises from the eligibility of a linguistic or semantic manoeuvre, rather than as a physical or metaphysical feature of the world.

Schiffer uses the example of 'being a dog' as an example. We can endorse the statement 'Lassie is a dog' and from this sentence we endorse 'Lassie has the property of being a dog'. We move from a sentence containing only one singular term 'Lassie' to an equivalent sentence containing the new singular term 'the property of being a dog'. This new singular term refers to pleonastic property of being a dog. We have introduced it by a something-from-nothing transformation: by moving from a sentence where no property is referred to, to an equivalent sentence which refers to one. Whether being a dog is a pleonastic property may be contentious, but I believe this fits what the deflationist means when they claim that their property of truth is insubstantial.

Deflationists are happy to endorse sentences of the form "'s" is true' where the only singular term is 's'. Those who endorse a property, are happy to licence the linguistic move from these sentences to those of the form "'s" has the property of being true'. This now introduces the singular term 'the property of being true' which refers to the truth property. For deflationists, this property exists in virtue of this linguistic transformation and has no substance or role beyond this. Their truth property is claimed to have no nature to discover, and no use beyond that which can be described the logical-linguistic-semantic behaviour of the word 'true'. This fits with a pleonastic description of the property, whereby this property has no existence or role beyond a certain something-from-nothing transformation from a sentence containing the word 'true'.

This is in contrast to substantive theories of truth, which posit a property of truth with existence that consists in more than an appropriate language game. For instance, in correspondence with facts, coherence with a body of truthbearers or epistemic success. The natures of these truth properties do not obtain from something-from-nothing transformations, but by introducing metaphysical or epistemic notions and appropriate relations between them. For an inflationary theory of truth, that s has the property of being true tells us more than just 's' is true, it tells us that 's' corresponds to facts, or coheres with other truthbearers, or that knowing s leads to epistemic success, etc. These features do not come from an appropriate language game using the word 'true', but from the property of truth itself.

This understanding of the deflationist truth property as a pleonastic property is not unique, but does appear to have attracted remarkably little discussion. Crane (2013, §3.4), for instance, identifies a minimal truth property as an example of a pleonastic property, since the minimalist only identifies truth as a property in the sense that it is what we predicate of a sentence when we say that it is true. Künne (2003, p. 89) also appears to endorse a pleonastic view of the truth property, as part of his deflationism, although not in these terms:

> The predicate 'is true', I have argued, is a genuine predicate, hence truth is a *property* under that prodigial reading under which whatever is ascribable by a genuine predicate is a property. In so arguing, one does *not* incur a commitment to a 'realist' view of such properties.

Lynch (2009, p. 106-7), in a discussion of deflationists' metaphysical views, similarly endorses a pleonastic criterion of a deflationary truth property:

> contemporary deflationists ... allow that the truth concept does express a property – in the same sense that the concepts of existence of identity express either a property or relation. Such properties, we might say, are *metaphysically transparent* or pleonastic properties.

Given this quote, and the way I presented pleonastic properties earlier, one may have the conception that a transparent (revelatory) property and a pleonastic property are equivalent. It is the case that any pleonastic property will be revelatory, since pleonastic properties can only be discovered by an appropriate something-from-nothing transformation. To grasp the property, we need only grasp the relevant concept and make the appropriate predicate to property transformation. On the other hand, not all revelatory properties need be pleonastic in nature (although it seems that the vast majority shall be). Consider the example of colours. We might consider colours to be revelatory. Johnston (1992), for instance, endorses this position, since grasping the full concept of blue might fully reveal the property. On the other hand, granting that blue is a revelatory property, it may be denied that our grasp of the property of blue comes from an appropriate something-from-nothing transformation. Instead, it seems that we discover the property by direct visual perception, rather than a language game. This shows that Damnjanovic's notion of a deflationary truth property as revelatory is not incorrect, but does not characterise deflationary truth properties, and instead that this epistemic view of deflationism follows from an appropriate metaphysical conception of deflationism.

This notion of a deflationary truth property as a pleonastic truth property is an appropriate foundation for both Edwards' and Wyatt's conceptions of deflationism as well. For the deflationist wishing to make use of Lewis' sparse/abundant property distinction, it is natural to view the pleonastic properties as falling on the abundant end of the spectrum. These properties are not natural or fundamental to the world, and instead only exist in the sense that we can introduce them linguistically. Similarly, this allows us to spell out the notion of 'consists' that Wyatt makes use of. What can be understood by 'consists' here is a theory of the property that does not depend upon a predicate that expresses it. The pleonastic properties allow no constitution theory other than a theory of the predicate that expresses them, since this is all there is to them. This is in contrast to genuinely substantive properties, which have constitution theories which have no dependence upon the predicate: for example the property of water consists of H_2O molecules, but does not consist in being expressed by the predicate 'is water'. I find that the conception of a deflationary truth property as a pleonastic property is a useful one which can explain and clarify, rather than refute, other conceptions of the insubstantiality of truth.

3.5 Conclusion

In the previous section I argued that we should view deflationism about truth as a logical-linguistic-semantic theory of the word 'true' and the conception that this is all that is needed of a theory of truth. This leads to the notion of a deflationary truth property, if the theory admits that the word 'true' functions as a predicate, as a pleonastic property. These are properties which only exist in the sense that they result from a linguistic 'something-from-nothing' transformation. This notion of a deflationary property adequately explains why a deflationary truth property is revelatory in nature and can be used to clarify what it would mean for a deflationary truth property to be abundant or lack a constitution theory.

This conception of deflationism is far wider than solely a T-Schema, which as I argued in Section 3.2 does not adequately capture deflationism about truth. This has real philosophical consequences. If I am correct, then many arguments against deflationism about truth in the literature are, in actuality, only arguments against the T-Schema being adequate as a theory of truth. For example, Liggins (2016) recently argues that there is no adequate deflationary explanation of the

explanatory asymmetry of truth – that P explains that 'P' is true, rather than the other way round. Yet Liggins' arguments interpret deflationism as a theory of truth based solely upon a T-Schema, and if I am correct then Liggins has only argued against a collection of deflationary theories of truth, and has no argument against deflationism as a whole. Greenough (2010) also challenges deflationism about truth, by arguing that it is incompatible with truth-value gaps. Again, this argument relies upon the understanding that deflationism about truth is characterised and exhausted by some kind of T-Schema. Even if these arguments are knockdown refutations of any form of T-Schema as a sole theory of truth, they do not, as currently stated, significantly challenge deflationism about truth more widely.

By understanding deflationism about truth as wider than a T-Schema, we find that arguments against deflationism are, in actuality, only arguments against specific deflationary theories of truth. I have, however, given a positive proposal for what deflationism about truth is, and this conception can be challenged. For example, Lynch (2009) argues that deflationism about truth is incompatible with the 'truism' that it is correct to believe 'P' if and only if 'P' is true. Lynch phrases his challenge by arguing that a T-Schema is insufficient to explain this. It seems plausible that Lynch would regard his argument as extending to any logical-linguistic-semantic theory of the word 'true'. His basic criticism is that this normative aspect of truth cannot be reduced to a purely descriptive character, such as a logical-linguistic-semantic theory. Deflationism about truth is still open to criticism on my conception, but requires subtler arguments than showing a T-Schema is inadequate to perform a certain role.

This also opens routes to explore new deflationary theories of truth, with potentially useful applications. For example, often deflationary theories are viewed as inadequate in light of the semantic paradoxes. Beall and Armour-Garb (2005) suggest that these are particularly problematic for deflationary theories of truth and Simmons (2018) recently argues that his solution to the semantic paradoxes is incompatible with deflationism. Each author identifies deflationism with a T-Schema theory. Allowing deflationary theories to make use of further logical-linguistic-semantic resources lets potential solutions and compatibility between current solutions and deflationism as a whole emerge. Such remarks may appear optimistic, but I shall aim to show that at least one deflationary theory of truth can do this formally in Chapter 5, Section 5.2.5.

Before we can see whether deflationism can tackle the paradoxes formally, and to tackle my wider question of whether formal theories of truth support or oppose deflationism about truth, we need to understand which formal theories of truth are deflationary. This will be the main question of Chapter 4 and will make use of my conception of deflationism proposed in this chapter – as a logical-linguistic-semantic theory of the word 'true' . This will also allow us to answer the dilemma proposed at the end of Chapter 2, whether deflationists should endorse or oppose a conservative theory of truth. I will argue that deflationists should oppose conservativity and that we should see all current axiomatic theories of truth as deflationary theories of truth.

Chapter 4

Deflation, Formalisation and their Intersection

At the end of the previous chapter, I submitted that a deflationary theory of truth is a logical-linguistic-semantic theory of the word 'true'. In this Chapter I will use this proposal to analyse which formal theories of truth are deflationary theories of truth. This is key to answering whether research in formal truth theory supports or opposes deflationism and to answer the question at the end of Chapter 2 – whether deflationists should accept or reject conservativity. I will argue in this chapter that deflationists should reject conservativity and that, due to my conception of deflationism, all current axiomatic theories of truth are deflationary. With this understanding in place, it will be argued in Chapters 5 and 6 that formal truth theories support deflationism about truth.

Chapter Abstract

I question which formal theories of truth are deflationary theories of truth and argue that all axiomatic theories of truth are deflationary theories of truth. I examine the proposed formal criteria of deflationism in the literature: proof-theoretic conservativity, model-theoretic conservativity and logicality, and argue that each is inadequate. The criterion of proof-theoretic conservativity conflates deductive power with explanatory power and the criterion of model-theoretic conservativity relies upon a problematic formal understanding of insubstantiality. Logicality fairs better, but deflationism should be understood as wider than logicality, and hence tests of logicality can never rule out a formal theory of truth as deflationary. I argue that all axiomatic theories of truth are logical-linguistic-semantic in nature and

thus, using my understanding of deflationism in Chapter 3, these should all be regarded as deflationary.

4.1 Introduction

Traditionally it is claimed that truth is an important philosophical notion, capable of providing explanation to diverse and important areas of philosophical exploration such as knowledge, logic and science. Knowledge is justified true belief; the meaning of connectives are given by their truth conditions; science aims towards the truth. Truth plays an important explanatory role in all these areas of philosophy, and more.

Deflationists about truth step in and disagree. The truth predicate may play an important expressive role in these statements, but this is a mere linguistic convenience. The truth predicate cannot play a serious explanatory role, so claims the deflationist, and instead is a device of 'semantic descent' and 'semantic ascent' which enables one to move from the expression of a sentence as true to the content of the sentence itself and back again. Truth is an insubstantial notion without metaphysical depth.

Whether every use of the truth predicate in natural language is one of semantic descent or ascent, not of a more serious explanatory nature, is not easy to prove. Logicians have therefore taken this question into the formal domain, where formal methods can be brought to bear upon the issue. This gives rise to the so-called 'conservativity arguments' against deflationism.

Two such arguments have been brought against the deflationist. The first is the argument from proof-theoretic conservativity: that a deflationary theory of truth ought not to be able to prove more than its background theory, or else it is capable of providing explanation. The second is the argument from model-theoretic conservativity: that every model of the background theory ought to be able to be expanded to a model with an interpretation of the deflationary theory of truth, else it provides substantive metaphysical content. These arguments then claim that no good theory of truth satisfies either proof-theoretic or model-theoretic conservativity, and thus deflationism about truth must be incorrect.

This is not the only way that logicians have explored whether deflationism about truth is correct. More recently, there have been advocates[1] for the position

[1]Künne (2003), McGinn (2000), Damnjanovic (2010), Horsten (2011) and Galinon (2015)

that deflationism about truth claims that truth is (at least something like) a logical property. This can be analysed in terms of proof-theoretic conditions of logicality, such as whether truth can be given by inference rules, or semantic conditions of logicality, often given by invariance under certain functions. These arguments, *contra* the conservativity arguments, often find that formal truth properties are logical, and thus deflationism about truth is correct.

This shows that at least some error is being made here, between advocates of conservativity on the one hand and the advocates of logicality on the other. Clearly both camps cannot be correct. In this chapter I shall look critically at these arguments to conclude that none of these are adequate arguments for or against deflationism about truth. I argue instead that deflationism, properly construed, sees that *all* axiomatic theories of truth are deflationary, and thus the success of deflationism about truth rises or falls with the success of finding a suitably adequate axiomatic theory of truth.

4.2 The Case for and against Proof-Theoretic Conservativity

The first argument connecting deflationism about truth and formal truth theory comes from the argument for proof-theoretic conservativity, mentioned in Chapter 2. The conservativity argument claims that a deflationary truth theory ought not prove more than its background theory, since otherwise the truth predicate is serving an important explanatory function to the theory. Since even basic truth theories are not conservative over arithmetic, the critics conclude that thus the deflationist position is flawed.

Horsten (1995), Shapiro (1998) and Ketland (1999) have all argued that a deflationary theory of truth ought to be (proof-theoretically) conservative over its base theory. Informally, this means that the theory of truth cannot prove anything that the base theory cannot prove. More formally, this is stated:

Definition 4.2.1. *Let $\mathscr{L}_B$ be a language and B be a theory in this language. A theory of truth T for B is proof-theoretically conservative over B if whenever $T + B \vdash \sigma$, for some sentence σ in $\mathscr{L}_B$, then $B \vdash \sigma$.*

have all made claims similar to this.

The advocates argue for this using deflationists' claims that truth is not a serious explanatory notion. Were truth to be a serious explanatory resource, then one could derive new consequences from it. If truth is merely an expressive resource, without any explanatory power, then one should not be able to derive entirely new consequences just by adding a truth predicate. Thus if any reasonable theory of truth allows one to derive entirely new consequences, then truth must be an explanatory notion, and hence deflationism about truth is incorrect.

This argument, if sound, is a strong challenge to the deflationist. Consider the test case where $\mathscr{L}_B$ is the language of arithmetic $\mathscr{L}_A$ and B is the first order theory of Peano Arithmetic (PA). Even relatively simple theories of truth can prove the consistency of PA. An example of one such theory was shown in Chapter 2, the theory of typed compositional truth without induction (CT^-), closed under an 'extended T-schema'.[2] The consistency of PA is not provable in PA alone due to Gödel's Second Incompleteness Theorem. Thus, taking natural language as containing arithmetic and assuming that the truth predicate satisfies these simple properties, we already see non-conservativity phenomena. The argument claims that the truth predicate has added substantial semantic consequences, which were not present within the original theory. Truth plays an explanatory role, therefore, and the deflationist position as stated is untenable.

The conservativity argument has prompted a variety of interesting responses, which provide escape from the argument using both philosophical and formal strategies. Some of the more successful strategies, in my eyes at least, are Field's (1999), Nicolai's (2015), Horsten and Leigh's (2017) and Fujimoto's (2019).

Field's (1999) response is to argue that a distinction between the content of the truth predicate and the arithmetical content of the theory needs to be drawn. The Tarskian compositional clauses for truth (CT^-) alone are conservative over PA. Thus, the deflationist can hold a conservative truth theory. If this theory is introduced over arithmetic, however, then due to the indefinite extensibility of induction we also have the additional mathematical content of inductive axioms for the language with the truth predicate. It is these mathematical induction axioms that provide substantive explanatory power, rather than the truth predicate which alone is conservative. My results in Chapter 2 argue against this strategy, however, and show that ensuring CT^- satisfies an extended T-schema for nonstandard mod-

[2]The details of CT^- and this extended T-schema are provided by Definition 2.3.1.1 and Definition 2.5.2 respectively.

els of arithmetic results in a nonconservative theory of truth, without requiring additional mathematical content for the theory. Cieśliński (2007) provides alternative results[3] arguing against Field by showing that a compositional truth predicate satisfying natural logical properties, not mathematical, is also non-conservative.

Nicolai's (2015) strategy is to draw a distinction between the domain of objects and the domain of syntax. Ordinarily, when working with a formal theory of truth, the truth predicate applies to Gödel codes of sentences, numbers which code sentences of $\mathscr{L}_A$. Nicolai argues that we should make a distinction between numbers and sentences, which ordinarily is not done. When this is made formal, with a 'disentangled' theory of syntax, the resulting theory is conservative over arithmetic.

Horsten and Leigh (2017) argue that it is our implicit commitment to reflection principles which result in non-conservativity, and not our theory of truth itself. The basic theory of truth TB consisting of Tarski Biconditionals of the form $Tr(\ulcorner\sigma\urcorner) \leftrightarrow \sigma$, for σ not containing the truth predicate, is conservative over arithmetic. Once we formally reflect upon this theory, by adding reflection principles such as $\forall x[Prov(\ulcorner\varphi(\dot{x})\urcorner) \rightarrow \varphi(x)]$ for each φ in the language with the truth predicate, we reach nonconservative theories of truth. They argue that our basic concept of truth TB is conservative, and it is by adding in our implicit commitment to these reflection principles that we get nonconservativity results.

Fujimoto's (2019) response is different, and he argues that the case should not be examined over arithmetic at all, but instead set theory. The deflationist, when adding a theory of truth to mathematics, is not interested in a weak subtheory, but instead the entirety of its rich and varied landscape. Considering ZFC as our background theory, rather than PA, he proves that Tarskian compositional clauses for truth with full induction (CT) are conservative over ZFC. This shows that care must be taken when choosing the base theory to which we add a theory of truth, and that arithmetic is something of a 'red-herring'.

All of these strategies are interesting valid responses that a deflationist can make to avoid the force of the conservativity argument, but I propose that, whilst interesting results in and of themselves, they are not needed by the deflationist. These responses all implicitly accept the argument's claim that a deflationary theory of truth ought to be conservative, and then show ways the deflationist can

[3]These results rely upon a proof by Wcisło and Łełyk (2017) which fixes a previous error in one of Cieśliński's cited results.

argue that their theory is conservative. Yet outside the formal domain, one would be hard-pressed to find a deflationary philosopher of truth who accepts that their theory of truth ought to be conservative. Why is this the case?

The conservativity argument claims that the truth predicate should not have deductive power, since this is a formal explication of the thesis that truth is not a powerful explanatory resource. There is an important distinction to be made here though, identified by Cieśliński (2015), between deductive power and explanatory power.

The deflationist about truth claims that truth is not an explanatory notion, but this should be distinguished from the claim that truth cannot be used for (truth-free) deductions. Conservativity formalises this latter claim, that truth cannot be used for deductive purposes, and says nothing of whether these deductions constitute explanations or not.

Cieśliński draws a distinction between two types of proof. All proofs offer an argument for the veracity of a claim. From a proof of σ, one infers that σ is true, they are justificatory. Yet, other proofs go beyond this, and allow one to explain why it is, say, that $\exists x\varphi(x)$ holds, rather than $\neg\exists x\varphi(x)$. Consider proofs which are directly constructive, and give an a such that $\varphi(a)$, against proofs which merely prove why $\exists x\varphi(x)$ has probability 1, without any example of an a.[4] The former provide a substantive explanation that $\exists x\varphi(x)$ because this constructed object a has the property φ. The latter, indicates that $\exists x\varphi(x)$ is true, without (perhaps) an explanation of why this is the case.

In my terminology, we should accept that a theory of truth can offer deductive power, without this translating to explanatory power. This distinction between explanation and deduction is clear in a general setting. I may deduce that it is raining based on looking outside the window, but an explanation of why it is raining would involve some further meteorological facts and relevant evidence of clouds, humidity, etc. We make deductions about the world all the time, even if we cannot always offer an explanation of why they are the case. I think sense perception is a clear case of this, where perception of an event, such as an eclipse, is enough to deduce it's occurrence, without any clear explanation for why it occurs.

In a purely mathematical setting, this distinction is less clear. As Cieśliński notes, the cases of when a proof is explanatory versus merely justificatory are not

[4]Alon and Spencer (2016) provides a detailed summary of this proof method and its wide-ranging and powerful use in modern combinatorics.

yet well understood in philosophy of mathematics.[5] Whilst this distinction can be made use of by the deflationist, it alone is not enough to deny the conservativity argument. There are two possibilities, even given this distinction, in which the conservativity argument is still valid. The first possibility is that some proofs are explanatory, and some are merely deductive, but that the use of truth predicates in non-conservative proofs are explanatory. For a response to this possibility, I refer the reader to Cieśliński's (2015, p. 79-81) argument, who carefully analyses the use of theories of truth in proofs of non-conservative results. I would instead like to focus on the second possibility, that whilst there is a general distinction between deductions and explanations, this does not occur within a formal framework, and in the formal domain all deductions are explanations.

I am sympathetic to this view. In some sense it seems correct that all proofs are explanatory. Given that $B \vdash \sigma$ we get a clear explanation that from the axioms of B and the inference rules of logic one can explain that σ is true via valid argumentation. What would an explanation consist in, if not some true/assumed statements and a sequence of agreed rules flowing from these assumed statements to the conclusion? Suppose our probabilistic proof shows that the probability $\neg\varphi(x)$ holds for an arbitrary x is strictly less than 1. This means that there must be some x for which $\neg\varphi(x)$ does not hold, i.e. it satisfies $\varphi(x)$. One arrives at an explanation for why $\exists x\varphi(x)$ is the case from the axioms of probability and the law of excluded middle, even if there is no explanation (or justification) for any specific a that satisfies $\varphi(a)$. It appears to be this aspect of proofs that adherents of the conservative argument are remarking upon, that a proof will always provide some explanation why σ is true. If $B + T \vdash \sigma$, and B alone could not, then by virtue of being a proof we have an explanation of σ (relative to $B+T$), which is not present in B. This distinction between deduction and explanation, in this technical setting, might well be critiqued as collapsing altogether. Sense perceptions may offer a case against deductions and explanations being equivalent in all domains, but in the formal domain we have no sense perceptions available to draw a distinction between them so neatly.

This distinction between deduction and explanation is important within even a formal framework, however, and to argue otherwise is to ignore important meta-theoretic concerns about B from our reasoning. The aim of the conservativity

[5]Mancosu (2015, §4) provides an overview of notions of explanations within mathematics, but concludes "work in analytic philosophy in this area has only just begun" (Mancosu, 2015, §7).

argument is to take a toy model of natural language (arithmetic, usually) and apply formal methods to it. We make certain assumptions about this toy model, however, such as sufficient mathematical strength to talk about its own syntax, along with consistency and sufficient expressibility.

Suppose, purely for example, that unbeknownst to us ZFC is inconsistent. Take two set-theorists A and B. A proves the novel theorem $\mathrm{ZFC} \vdash \sigma$, to much rejoicing, for a deduction, and it is (too) quickly concluded explanation, of σ has been found. In fact, A later improves on her result and shows that one can carry out the same proof in a much weaker consistent base theory (say PA). This proof follows much the same reasoning, making use of some clever coding tricks, and it is concluded we most definitely have a deduction of, and explanation of, σ. Unfortunately, B later shows that $\mathrm{ZFC} \vdash \bot$, which gives B a rather trivial proof of A's first result: $\mathrm{ZFC} \vdash \bot, \bot \to \sigma$ and hence $\mathrm{ZFC} \vdash \sigma$. This second ($B$'s) proof should certainly not count as an explanation of the truth of σ. It is a formal deduction of σ from ZFC, but is not in any way an explanation of σ. The first proof (A's), on the other hand, at least when ran through PA, certainly constitutes an explanation of σ.

Explanation, at the very least, assumes consistency within the background assumptions. Deduction, on the other hand, does not. Inconsistent theories still allow us to make deductions, it just so happens that unless we go paraconsistent they allow us to deduce everything. Explanation thus must go beyond deduction, even in the formal domain, as it depends upon these metatheoretic considerations. The conservativity argument claims that a deflationary truth theory should not allow one to make novel deductions, because a deduction of a sentence is an explanation of the sentence, but this cannot be the case without these metatheoretic considerations. It is these metatheoretic assumptions which highlight the disconnect between natural language and our toy model. Metatheoretic assumptions beyond natural language are not possible within natural language itself, whereas in our toy models of interest, such as those involved in truth theory, it appears required. The conservativity argument approximates explanation with deduction, but in at least this one important aspect, it is not an accurate simulation.

With this distinction between deduction and explanation clear, even in the formal domain, we can question whether deflationists claim, or are committed to claiming, that truth is not deductively powerful. I believe that this is not the case, and the way we use the truth predicate in everyday language allows us to form deductions with it. Truth is a deductive notion for natural language. This holds

just as much for the deflationist, as other theorists of truth, and does not contradict the deflationist's stance that truth is insubstantive and non-explanatory. Deflationists should not claim that the truth predicate cannot be used for deductions, but instead should, and do, make the much weaker claim that truth cannot be utilised for explanation.

For example, consider the following arguments one might find, all making use of the truth predicate, and all of which have 'truth-free' conclusions:[6]

In theology one might make the following deduction:

P) Everything in the Bible is true

P) The bible says that murder is wrong

C) Therefore, murder is wrong.

In law, one might deduce:

P) Everything the witness said is true

P) The witness said that he committed the crime

C) Therefore, he committed the crime

Even in science, one would say:

P) To the best of our knowledge, the standard model of cosmology is true

P) The standard model of cosmology implies dark energy exists

C) Therefore, to the best of our knowledge, dark energy exists

All of these deductions are valid and make use of the truth predicate. The truth predicate allows these deductions to be formed and it is hard to see how to finitely reformulate the premises of the arguments without implicit use of the truth predicate. The deflationist is happy that the first premise of each argument can be reformulated without the truth predicate, but the standard approach is to replace it with an infinite schema of premises of the kind:

P(σ) If 'σ' is in the Bible, then σ.

Q(σ) If 'σ' is something the witness said, then σ.

R(σ) If 'σ' is a consequence of the standard model of cosmology, then σ.

[6]Fujimoto (2019) provides a similar example of using the truth predicate in natural language to make a deduction, but one in which the conclusion also involves the truth predicate.

The argument produced from this is now infinitely long in length, however, and thus no longer a valid deduction classically without recourse to infinitary logics. The deflationist has no problem with this, and specifically argues that one use of the truth predicate is in reformulating infinitely long conjunctions as a single finite sentence. The deflationist does not want to reject that truth cannot play a deductive role, since in these examples it clearly does so.

This deductive power of truth cannot be avoided by replacing the first premise of each argument with only finitely many instances of these schemas, either. This ignores the counterfactual statements implied by the first premise. Consider a restating of the second argument, involving two police detectives talking to a newsreporter. Detective Carter tells the reporter that everything the witness says is true, for the witness is extremely honest and would never lie under oath, although Carter did not turn up to court and does not know what the witness said. The reporter then speaks to Detective Lee and learns that the witness said the defendant committed the crime. Thus, she deduces that the defendant did indeed commit the crime. Carter's statement cannot be reformulated into a finite truth-free statement, since his statement is intended to cover everything the witness could possibly have said in court, including all the infinitely-many statements that she did not.

In fact, one can form deductions in this way when both premises utilise the truth predicate. Consider a new court case, with another exceptionally truthful witness:

P) Everything the witness said in court is true
P) Everything said from 14:00-14:15 in court was false
C) Therefore, the witness was not talking from 14:00-14:15

In this example, it again is not possible to reformulate these premises to avoid the truth predicate and to retain the counterfactual instances that they imply. Truth is therefore doing deductive work over natural base theories (theology, law, and science) and thus we should not expect a deflationist to be committed to truth being conservative. Truth has deductive power and can be used to make new (truth-free) deductions. The deflationist is happy to accept this point, but can still hold the claim that in these cases the truth predicate is not providing

explanatory power, but merely a linguistic service.

This distinction is not one which is anathema to deflationists, but one that can be found implicitly endorsed. Consider Putnam's (1978) argument in which truth is professed to serve an explanatory role. Putnam has argued that we explain that a theory T has empirical success by virtue of referring to T's truth, e.g. 'A successful theory is one which is true', and thus truth helps explain the notion of a successful theory in philosophy of science. The deflationist Horwich (1998) responds that the role truth plays here is purely expressive, and not an explanatory one, and in fact Putnam's claim can be replaced by a schema: T is a succesful theory if and only if T, where we range over all T. To quote Horwich (1998, p. 49):

> "...No further explanatory depth is achieved by putting the matter in terms of truth. None the less, use of the truth predicate in this sort of context will often have a point. When it gives us certain economy of expression, and the capacity to make such explanatory claims even when we don't explicitly know what the theory is, or when we wish to generalise"

For the deflationist, by using its function of forming generalisations and allowing blind ascriptions, the truth predicate enables deductions to be made, yet this does not entail that truth is providing an explanatory role within any of these deductions.

We can see this in the arguments above, where truth is providing a deductive capacity. In the third example in particular, for example, the argument certainly does not seem to provide adequate explanation as to why dark energy exists, and this is certainly not given by the role the truth predicate plays. An explanation for this would consist in numerous facts of cosmology and how they entail that dark energy exists - that one can deduce the conclusion using the truth predicate does not serve as an explanation of the claim. Deduction and explanation are not equivalent, and the deflationist can happily endorse that truth provides deductive power, whilst rejecting explanatory power.

Putting this together, we find that the deflationist has no quarrel with truth playing a deductive role for theories and that some formal proofs are purely deductive, not explanatory. The conservativity argument, hence, no longer stands up. The standard conservativity arguments take Tr as some adequate theory of

truth, then show that PA + Tr $\vdash$ *Con*(PA), and since by Gödel's second incompleteness theorem we know PA $\nvdash$ *Con*(PA) we get that PA + Tr is non-conservative over PA and thus Tr is doing explanatory work.

It would be quite strange to contend that Tr is doing explanatory, rather than deductive, work here however, once this distinction is emphasised. The consistency of PA is an already-present metatheoretic commitment within the background, since if PA were not consistent, then it could prove everything, and in particular it would already prove *Con*(PA) and Tr would be conservative over it. That Tr allows us to formally deduce it, does not constitute an explanation of this fact, since the commitment to *Con*(PA) has already been accepted.

If one were truly sceptical of the consistency of PA and were hoping for an explanation of it, then providing a formal proof of it in PA + Tr would offer no such explanation, since this would be provided whether PA was consistent or not. One has to already accept the consistency of PA to accept this as a deduction of relevance.

Does this mean that non-conservativity results of truth theories cannot tell us anything? I do not believe that this is the case. Whilst showing that PA + Tr $\vdash$ *Con*(PA) does not offer an example of the truth predicate providing substantive explanatory power, this does not mean that no non-conservativity result could do this. Consider the extreme example where one shows that PA + Tr has sufficient mathematical strength to interpret ZFC+ many Large Cardinal Axioms. This (presumably) unlikely result seems that it would go beyond a merely deductive inference and instead show that truth can play an important explanatory role in mathematics, in taking one from arithmetic to strong set theory. Something like this result would put the deflationist in a weaker position, in my eyes.

No proof of this has been offered, however, or it appears currently is likely to be offered. So far, instances of non-conservativity that have been offered are of the consistency kind and seem to fit comfortably within the deductive/explanatory distinction. It therefore appears that deflationists should not be worried by the conservativity argument and results of the non-conservativity of strong truth theories. I thus take proof-theoretic conservativity to be inadequate as a method which formally distinguishes the deflationary theories of truth from the non-deflationary theories.

This answers the question left at the end of Chapter 2. Deflationists should not endorse conservativity, and thus can endorse the desirable proof-theoretic features

of an extended T-Schema.

4.3 The Case for and against Model-Theoretic Conservativity

If proof-theoretic conservativity is not the best decider of which formal theories of truth are or are not deflationary, then is there a suitable rival to take its place? Recently interest has surfaced in a different kind of conservativity: model-theoretic conservativity. Strollo (2013) has advocated that truth should be semantically, rather than syntactically, conservative over its base theory.[7]

The model-theoretic notion of conservativity instead looks at which models of a base theory can be expanded to models of the base theory and the theory of truth. If not all models of the base theory can be suitably expanded, then the theory of truth is not model-theoretically conservative. More formally, this is written as:

Definition 4.3.1. *We say that a theory $T \supseteq B$ is model-theoretically conservative over B if and only if every model of B can be expanded to a model of T.*

This is a model-theoretic definition which implies syntactic conservatvitity, but is not implied by it, as the following result states.

Proposition 4.3.2. *Let B and T be given as above. If T is model-theoretically conservative over B, then T is proof-theoretically conservative over B. There are theories B and T where T is proof-theoretically conservative over B, but T is not model-theoretically conservative over B.*

There are many examples which show that this converse does not hold, but perhaps the most appropriate example would be from the literature on truth.

Theorem 4.3.3 (Kotlarski et al. (1981))**.** *The theory of truth consisting solely of Tarskian compositional clauses* CT^- *is proof-theoretically conservative over* PA*, but only the standard model and recursively saturated models of* PA *can be expanded to a model of* $\mathrm{PA} + \mathrm{CT}^-$*.*

[7] Thus far support and criticism of this view appears to be lacking. Fischer and Horsten (2015) express interest in the notion of semantic conservativity, but do not go so far as to advocate it explicitly. Cieśliński (2015) is more critical of the view and argues against the appropriateness of semantic conservativity to deflationists for two reasons: firstly, there is no sound textual endorsement for this within deflationists' writings, and secondly, deflationists' preference for axiomatic accounts, not an interpretation within a model.

Given that I have already argued that proof-theoretic conservativity should not be prescribed to the deflationist, it might be thought that given these results model-theoretic conservativity is immediately too strong to prescribe also. I would not say that this argument follows quite so easily, however. These are different criteria, with different theoretical underpinnings, and the arguments should be evaluated on their own merit.

What are these arugments? Strollo has put forward two main arguments for this view. The first argument plays upon deflationary claims that the property of truth lacks 'substance'. The argument is that if not all models of the base theory can be expanded to models of the base theory with a property of truth, then truth is imparting substantive structure to the theory. Since truth is imparting metaphysical structure and specifying how the models of the theory should be, the truth predicate hence has important metaphysical 'weight' to it. One of the core claims of deflationism is to deny the 'metaphysical weightiness' of truth, and hence deflationists about truth cannot have a semantically non-conservative theory of truth (Strollo, 2013). This argument is then, similarly to the argument from proof-theoretic conservativity, supplemented with the additional premise that almost all theories of truth are not model-theoretically conservative over arithmetic, and thus deflationism about truth is incorrect.

Strollo's (2014) second argument adds specification to this first argument. His argument makes use of Edwards' (2013a) criterion of deflationism that Lewis' (1986) conception of sparse and abundant properties can distinguish between inflationary and deflationary properties of truth.[8] Under this understanding, a sparse truth property is inflationary, whereas an abundant truth property is deflationary. Strollo then argues that we should understand abundant properties (such as deflationary truth properties) as those which are semantically conservative, and the fundamental properties as not semantically conservative.

I immediately reject this second argument. I have criticised Edwards' criterion of deflationism in Chapter 3 Section 3.3, but even granting this, Strollo's argument does not appear to stand up. Consider the property of being a model of arithmetic with cardinality ω_1. This should not be regarded as a natural fundamental property of arithmetic at all, and thus should come out as an unnatural 'abundant' property, but it is not semantically conservative over arithmetic. Alternatively,

[8]See Chapter 3 Section 3.3 for more details, where I explore and critique this criterion of Edwards.

some genuinely natural/fundamental properties to arithmetic, such as exponentiation, are arithmetically definable and semantically conservative over models of arithmetic. It appears that there is a large disconnect between the notions of sparse/abundance and (model-theoretic) conservativity/nonconservativity and we should not make the identification of them that Strollo does.

This does not speak to Strollo's first argument for model-theoretic conservativity, however, although it does mean that important details are missing. As with the adherents of proof-theoretic conservativity, Strollo runs this argument through models of Peano Arithmetic, but is this because all models of arithmetic are special in some way, or is it simply because the formal results have already been found for this theory? Whereas in the case of proof-theoretic conservativity, it might be highly reasonable to consider Peano Arithmetic for its mathematical power as a theory of arithmetic, it does not appear as reasonable to consider *every* model of Peano Arithmetic as equal, acceptable interpretations of the domain of our arithmetical reasoning.

Much of the philosophical interest in formal theories of truth is to explore theories which are able to approximate and simulate the behaviour of the truth predicate in natural language. Peano Arithmetic is a useful theory to work in for this because it is well-known, allows for arithmetisation of its own syntax via Gödel coding and, importantly, results on consistency are obtainable in a stronger, but still acceptable, metatheory (such as ZFC). It should be noted that we do not choose Peano Arithmetic because all of its models are regarded as on a par with one another, and are all equally good simulations of natural language.

Due to Löwenheim-Skolem theorems we know that there are models of all cardinalities of arithmetic. Consider a model of Peano Arithmetic whose domain has cardinality $\aleph_{29}$. This model does not have the same status as the natural numbers $\mathbb{N}$. It would be hard to find a number-theorist who regards the former as the natural domain of all the finite numbers! Similarly, we regard natural language as containing inherently finite expressions, and it would be hard to find a linguist who believes that there are finite expressions of length $\aleph_{28}$, but not of length $\aleph_{29}$. We should not regard all models as on a par, and just because some model cannot be expanded to a truth property does not mean that the model(s) of fundamental interest cannot be expanded to a truth property.

This is clear when we consider formalised provability predicates for PA, such as $Prov_{\mathrm{PA}}(\ulcorner\sigma\urcorner)$. All models of PA can be expanded to a model which inter-

prets this provability predicate, since it is arithmetically definable. If Strollo's model-theoretic argument is correct, we must conclude that this property is metaphysically insubstantial. Now, we consider the formalised provability predicate in conjunction with the principle $Prov_{\mathrm{PA}}(\ulcorner 0 = 1 \urcorner) \rightarrow 0 = 1$. Not all models of PA can be expanded to interpret a provability predicate and satisfy this further property, however. This is because it would enable one to prove *Con*(PA) which, again, by Gödel's Second Incompleteness Theorem is not syntactically (so therefore not semantically) conservative. As just stated, one of the reasons for interest in models of arithmetic is to prove results about their consistency. Following Strollo's argument, we are forced to conclude that the property '$Prov_{\mathrm{PA}}(\ulcorner 0 = 1 \urcorner) \rightarrow 0 = 1$' has substantive metaphysical substance, which certainly seems counter-intuitive. This principle is not positing new entities or a new type of entities, or any other metaphysical principle, but simply stating that if there is a proof within PA that $0 = 1$, then the theory should believe $0 = 1$. This addition seems metaphysically harmless, even if the logician knows it adds significant deductive (but perhaps not explanatory) content.

Non-conservativity in the model-theoretic sense does not always mean metaphysical strength, although it should be noted that I find it hard to infer what metaphysical strength could even mean in this context. It should be noted that in the same way we have metatheoretic commitments in exploring proof-theoretic conservativity, we have similar meta-theoretic commitments in exploring semantic-conservativity. In particular, much of the model-theoretic arguments are ran in the background of a stronger metatheory which picks out a definitive notion of the finite natural numbers, which only one model of arithmetic ($\mathbb{N}$) agrees with. Our metaphysics of a true model of arithmetic have been agreed upon long before truth comes into play.

It would be a strange move for a deflationist to object to a theory of truth being deflationary because it rules out certain unintended[9] nonstandard models of arithmetic. If a theory of truth rules out some nonstandard models of arithmetic, such as those which believe PA is inconsistent, then they might reasonably regard this as so much the better for the theory of truth, for they didn't believe that this model was true arithmetic anyway. On the other hand, a theory of truth which

[9]I emphasise the distinction here between ruling out *some* nonstandard models of arithmetic, against ruling all of them out. Whilst my arguments from Chapter 2 argue that we should not discount all nonstandard models of arithmetic, this does not mean the deflationist cannot reasonably ignore some of these.

rules out the metatheory's standard model $\mathbb{N}$ from consideration, such as FS, might reasonably be regarded as more objectionable. This surely is (at the very least one of) the intended model(s) of arimthetic, and thus it ought to be able to be expanded to a deflationary truth property, for deflationism to be correct. I leave it open whether this should be regarded as an aspect of deflationism, or simply an aspect of any adequate theory of truth, but my sentiments lean towards the latter.

Model-theoretic non-conservativity can be seen as ruling out some possible ways that the arithmetical world might be, in addition to what has already been stipulated by the base theory. At least some non-conservativity allows one to be more precise about suitable intentions and to rule out some alternative, undesirable, semantic comprehension of the theory. Deflationists about truth regard truth as an expressive notion, and one reading of this would be that they desire the truth property to narrow down the unintended interpretations of their domain of discourse. From this perspective, if truth was not semantically conservative, then it would have no real expressive power at all. It would not be able to tell us anything new or useful. It seems more likely for the deflationist to lean-into model theoretic non-conservativity phenomena, than shy away from it.

4.4 The Logicality of the Truth Predicate

Rejecting conservativity as a requirement of deflationism altogether thus seems like a reasonable way to proceed, but this leaves a gulf between formal theories of truth and the correctness or not of deflationism about truth. One way to bridge this gap is to consider deflationary claims that truth is something like a logical property.

This slogan can be found advocated by, or prescribed to, deflationists in various guises. Künne (2003, p. 338) writes that: "...Thus it appears reasonable to call truth a broadly logical property" and Field (1992, p. 322), writes that Horwich regards that truth "is a predicate of a very special kind, a logical predicate". A further example is that McGinn (2000) in the title of his book: *Logical Properties: Identity, Existence, Predication, Necessity, Truth* identifies truth as a logical property, and finally Damnjanovic (2010, p. 46) writes that "deflationists commit themselves to the idea that it [truth] is a logical predicate, and the concept of truth is a logical concept.".

Both Galinon and Horsten can be seen as taking this as a good way to connect

formal theories of truth and deflationism about truth. One can use formal notions of logicality to test whether formal theories of truth can be regarded as deflationary or not.

Horsten (2011) argues for a version of deflationism called inferential deflationism, where the truth predicate cannot be governed by general laws, and instead is essentially given by inference rules. Correspondingly, those formal theories of truth (such as PKF) which are given essentially by rules of inference are deflationary, whereas those which entail general principles about truth (such as CT and FS) are not inferentially deflationary. If truth has a substantive essence, then it appears that it would have general principles which govern it, if truth has none, then it must be an essentially inferential notion.

Horsten can be read as identifying formal theories of truth given by inference rules with inferential deflationism, and formal theories of truth which provide generalised principles about truth as not (inferentially) deflationary. This fits into views such as Prawtiz's (2006) that the meaning of logical constants is given by their inference rules (the introduction and elimination rules of natural deduction) and is an explication, as much as a justification, of the claim that truth is a logical notion.

This approach appears not to generalise beyond inferential deflationism however. It would be odd to reject that any deflationist can accept that truth satisfies some general principles, for many of them do so. Quine (1986) in elaborating his theory of truth defines it in terms of satisfaction and proceeds to define satisfaction in terms of the Tarskian inductive clauses. Quine (1986, p. 42) provides one general principle:

> For all sequences x and sentences y and y': x satisfies the conjunction of y and y' if and only if x satisfies y and x satisfies y'.

For Quine this a general principle about satisfaction (and hence truth), and further, one that can be formulated within the object language itself.

Künne (2003, p. 337) similarly endorses general principles about truth. His modest account of truth even takes the form of a general axiom:

$$\forall x(x \text{ is true } \leftrightarrow \exists p(x = \langle p \rangle \wedge p))$$

Both Quine and Künne are thought of as deflationists about truth, yet believe in and advocate general principles about truth. Horsten's proposal cannot be

generalised to see whether any formal theory of truth is deflationary, only whether those theories are inferentially deflationary. This means that, at best, we only have a formal notion of inferential deflationism. Whilst I am sympathetic to this connection between the two domains, I seek a general connection which holds for all deflationists, and thus an alternative proposal.

Galinon (2015) connects deflationism and logicality in a different way and argues that the debate should be focussed on the distinction between expressive and explanatory notions. Galinon observes that one obvious candidate for notions which are expressive, but not explanatory, are the logical notions, and in fact this lines up (as I also observe above) with what many deflationists claim truth to be. Galinon therefore argues that we should investigate whether formal theories of truth fit into various proof-theoretic and semantic frameworks for logicality, or not, as a good test to the veracity of deflationism about truth.[10] In the following section I explore one way that this investigation should go and contrast it to Bonnay and Galinon's (2018) own exploration. I use this comparison to show the flaws with connecting formal theories of truth and deflationism about truth currently using formal criterion of logicality.

4.4.1 The Non-Invariance of Truth

One positive route to explore Galinon's proposal would be to explore the relationship between truth and Tarski's (1986) semantic conception of the logical notions as those which are invariant under permutations of the world. In what follows I shall provide an example of how such an investigation could go, and show how it highlights issues with making an identification between logicality and deflationism.[11]

Tarski (1986) argues that the logical properties are those which are invariant under all permutations of the 'world', and the non-logical properties are those which are not. For convenience, I shall understand the 'world' to mean $\mathbb{N}$, although Tarski seemed to have something much broader in mind, a universal domain of everything. Since arithmetic is contained in the world, if truth is not a logical

[10] Galinon (2010) investigates the connection between formal theories of truth and proof-theoretic frameworks of logicality in his PhD thesis.

[11] It should be emphasised that Galinon (2015) does not argue that a deflationary truth predicate must be a logical notion, but offers this as a better avenue of exploration than proof-theoretic conservativity.

notion for $\mathbb{N}$, then it certainly won't be logical in the world. I will show that by taking the truth property as a materially adequate Tarskian predicate *Tr* such that $\mathbb{N} \vDash \sigma$ iff $(\mathbb{N}, Tr) \vDash Tr(\ulcorner \sigma \urcorner)$ we get that truth is not a logical notion, since it is not permutation invariant.

Tarski's criterion is a fairly simple classification of the logical properties that appears extensionally correct. For example, it finds that conjunction is a logical property. Understand $P(x) \wedge Q(x)$ to mean $\{x : P(x)\} \cap \{x : Q(x)\}$. Then for any permutation π of the domain we get that $\pi(P(x) \wedge Q(x)) = \pi(P(x)) \wedge \pi(Q(x))$. Contrast this to the property $x \sim y$ iff $2 \cdot x = y$. This is not a logical property and Tarski's criterion shows this. Consider the permutation π where $\pi(2) = 4, \pi(4) = 2$ and all other objects remain unchanged. Then we have that $2 \sim 4$ (since $2 \cdot 2 = 4$), but it is not the case that $\pi(2) \sim \pi(4)$ (since $2 \cdot 4 \neq 2$). This account extends to more 'substantial' properties such as blueness. If we consider a domain where some (but not all) of the objects are blue, then a permutation of a blue object with a non-blue object shows that this property is not invariant under permutation.

Applying Tarski's criterion to truth causes a problem for deflationists. This account finds that a materially adequate Tarskian truth predicate *Tr* is not a logical property, and this can be shown in two ways.

Proposition 4.4.1.1. *A materially adequate truth predicate Tr is not invariant under all permutations of* $\mathbb{N}$.

Proof 1. Consider $\ulcorner \forall x[x = x] \urcorner$ and $\ulcorner \forall x[x \neq x] \urcorner$ and a permutation $\pi : \mathbb{N} \to \mathbb{N}$ defined in the following manner:

$$\pi(x) = \begin{cases} \ulcorner \forall x[x \neq x] \urcorner & \text{if } x = \ulcorner \forall x[x = x] \urcorner \\ \ulcorner \forall x[x = x] \urcorner & \text{if } x = \ulcorner \forall x[x \neq x] \urcorner \\ x & \text{otherwise} \end{cases}$$

In this permutation we have that $Tr(\ulcorner \forall x[x = x] \urcorner)$ holds, but we do not have that $Tr(\pi(\ulcorner \forall x[x = x] \urcorner))$ holds as it is not the case that $Tr(\ulcorner \forall x[x \neq x] \urcorner)$ in the non-permuted model. The truth predicate is not invariant under all permutations.[12] □

This first proof shows that by treating sentences as objects these sentences

[12] An informal linguistic version of this proof can be found in Wyatt's (2016) *The Many (yet few) Faces of Deflationism.*

are able to be permuted in our domain. Analogously to considering the property of blueness, by permuting a true sentence with a false sentence, we get that the property is not invariant.

Proof 2. Consider $\ulcorner 2 + 2 = 4 \urcorner$ and $\ulcorner 4 + 4 = 2 \urcorner$ and a permutation $\pi : \mathbb{N} \to \mathbb{N}$ defined in the following manner:[13]

$$\pi(x) = \begin{cases} 4 & \textit{if } x = 2 \\ 2 & \textit{if } x = 4 \\ x & \textit{otherwise} \end{cases}$$

We then get that $\pi(Tr(x))$ is no longer a materially adequate truth predicate, and thus should not be regarded as a truth predicate at all. The reason for this is that $\pi(\mathbb{N}) \vDash 4 + 4 = 2$, but not that $Tr(\pi(\ulcorner 4 + 4 = 2 \urcorner))$. □

This second proof shows that the truth predicate is sensitive to how the model understands non-logical properties. Since these non-logical properties are not invariant under permutation (by definition), the truth predicate is unable to both remain accurate with regards to these properties and invariant under permutation.

It appears that by analysing even a very simple property of truth, we are forced to accept that truth is not a logical notion. This surely is far too quick a conclusion, however, and the deflationist instead should reject that truth is a logical notion in exactly the same sense as the logical constants are logical.

In fact, this is something that those deflationists about truth who clarify their conception of the logicality of truth frequently do, and something I argue against in Chapter 3 (Section 3.3). Horsten (2011) argues that truth is actually a logico-linguistic property, since the truth predicate is a predicate of linguistic objects and thus needs to be formed over a sufficiently rich base theory to formulate these, unlike true logical properties such as conjunction and quantification, which require no syntactic notions. Similarly, Künne (2003) clarifies that truth is only broadly logical, not purely logical, due to its reliance upon truthbearers. Deflationists can more reasonably be thought of as claiming that truth is a quasi-logical notion.

Tarski's motivation for taking invariance as the notion of logicality arose from Klein's 'erlangen project', where he presented different geometrical properties as those which are invariant under different criteria. If the purely logical notions are

[13] Without loss of generality we assume that $2, 4 \neq \ulcorner 2 + 2 = 4 \urcorner, \ulcorner 4 + 4 = 2 \urcorner$.

invariant under all permutations, and truth is only quasi-logical, then it may be thought to be invariant only under some permutations.

In fact, we cannot have a truth predicate which is invariant under any permutation (which isn't the identity permutation) and is materially adequate in the permuted model, as the next proposition states more formally.

Proposition 4.4.1.2. *Working in the language of arithmetic with a constant symbol for every number, $\mathscr{L}_A \cup \{c_n : n \in \mathbb{N}\}$, there is no non-trivial permutation $\pi \neq Id(x)$ such that $Tr = \pi(Tr)$ and Tr is materially adequate in $\pi(\mathbb{N})$.*

Proof. Suppose that π is a permutation such that $Tr = \pi(Tr)$, where π is not the identity permutation. Let $a \in \mathbb{N}$ be the least number which is permuted and denote b as the number such that $\pi(b) = a$. Consider the sentence $c_a < c_b$. We have that $\mathbb{N} \vDash c_a < c_b$, since b is permuted and a is the least number which is permuted. Therefore $\mathbb{N} \vDash Tr(\ulcorner c_a < c_b \urcorner)$ and thus $\mathbb{N} \vDash Tr(\pi(\ulcorner c_a < c_b \urcorner))$. We also know, however, that $\pi(\mathbb{N}) \vDash \pi(b) < \pi(a)$ and hence $\pi(\mathbb{N}) \vDash c_b < c_a$ and $\pi(\mathbb{N}) \nvDash c_a < c_b$. Thus $\pi(Tr)$ is not materially adequate for $\pi(\mathbb{N})$. □

Does this show, hence, that truth is not even a quasi-logical notion? Again, I think that this would be too hasty, for invariance of a property under some permutations seems like a poor conception of quasi-logical notions. We should not think of blue as a quasi-logical (and certainly not deflationary) property, as it is invariant under all those many permutations which only permute non-blue objects. Similarly, we should not think of water as a quasi-logical (and certainly not deflationary) property as it is invariant under all permutations which only permute objects not containing H_2O. Is there a better way to test for the quasi-logicality or not of truth properties, or is logicality also a flawed test for deflationism?

4.4.2 Against Logicality

Bonnay and Galinon (2018) have met the challenge to provide a way in which the deflationist can argue for the logicality of truth. They argue that it is unfair to apply the invariance test directly to truth, which is a property of sentences, since this (as I have shown) rather trivially shows that truth is not logical. For the deflationist, the truth precicate's utility is that it applies to sentences, to enable indirect talk of the world. By fulfilling its role as a device of semantic ascent and descent, the truth predicate cannot be a purely general property of objects, but

by necessity is a property of sentences, and thus it is unfair to apply invariance criteria directly.

Bonnay and Galinon's proposal instead analyses the logicality of truth by looking at the expressive power (definability of classes) that a truth predicate adds to an interpreted language. A somewhat brief overview of their approach is to take a generalised notion of invariance to be a similarity relation S[14] and look at the constants which are invariant under S, and hence logical under S. They then compare the classes definable by these constants with the classes definable by these constants and a materially adequate truth predicate. They show that if the logical notions under S are closed under definability, then the truth predicate can define no new classes. In other words, for a logic exactly generated by an S closed under definability, the truth predicate adds no expressive power.

Bonnay and Galinon offer this as a way for the deflationist to argue that truth is a logical notion. The truth predicate offers no expressive power that would not have already been expressible in principle by the logic in its most abstract general form. One problem with such an argument, however, is that many natural logics (and in particular first order logic) cannot be exactly generated by such an S. This tells us that over at least one natural standpoint for the logical notions, the truth predicate is actually not logical. This perhaps is not such a concern for Bonnay and Galinon - their proposal is not that truth is a logical notion under all conceptions of logicality, for that seems far too strong, but only under a given invariance notion of logicality which is closed under definability. Perhaps the conclusion to be drawn is that what is logical is not purely given by first order logic instead. I am concerned that there is a more general concern with logicality approaches to deflationism, however.

In my approach to logicality, I look at the truth property as a property of sentences. In Bonnay and Galinon's approach, they look at what such a truth property is able to express about the interpreted language. My approach ignores the semantic content of the truth property, to focus on its syntactic features, whereas Bonnay and Galinon ignore the syntactic features that the truth property has, to focus on the semantic expressions it can produce. Both approaches ignore something crucial about the truth predicate - that it has this feature of semantic

[14]For instance: Sher (2008) has argued the correct notion of invariance is isomorphism, Feferman (2010) has argued for strict homomorphism and Bonnay (2006) has argued for 'potential isomorphism'. S is meant to be neutral upon which of these is chosen.

ascent and descent that enables it to simultaneously be a property of sentences whilst making semantic claims about the domain. I believe it is this feature of truth that gives it a quasi-logical flavour. As Bonnay and Galinon show, the semantic content is 'logical', whereas I show the syntactic sentential property is 'non-logical'. Many deflationists, as I observe above, do not advocate for the position that the truth predicate is a logical one, for this reason. Deflationary truth should most properly be understood as something quasi-logical: logical in some guises, and non-logical in others.

With this understanding in place, identifying formal theories of truth as deflationary or not by their quasi-logicality or non-quasi-logicality appears as yet to be a prohibitive task. We do not have any tests for quasi-logicality, syntactic or semantic, and further at least some tests of logicality, such as Tarski's permutation invariance, appear very hard to adapt properly to classify the quasi-logical notions. It may be possible to formally capture such a property, but as of yet no such classification has been given. I have much sympathy for this as a positive proposal, but it is not clear to me what putting the matter in terms of logicality offers here yet, when formal tests of quasi-logicality are not readily available. This appears like an interesting avenue for further research, but one I shall not take up here, for I have an alternative proposal for which formal theories of truth are deflationary, based on my arguments in Chapter 3.

4.5 Deflating the Criteria of Deflation

I have a new proposal for connecting the debate between deflationism on one hand and the study of formal theories of truth on the other. My proposal is that we should regard all axiomatic theories of truth as deflationary, for these theories solely depend upon logical-linguistic-semantic (LLS) notions. This is based upon my arguments in Chapter 3 that the deflationary theories of truth are those which provide a theory of the word 'true' using logical-linguistic-semantic notions, rather than metaphysical, epistemic, or other concepts.

In Chapter 3 (Section 3.4) I argued that a deflationary theory of truth is a logical-linguistic-semantic theory of the word 'true'. I argued that deflationists provide a theory of the behaviour of the word 'true' using only logical, linguistic and semantic notions, and do not use further philosophical resources such as metaphysical or epistemic notions. For deflationists, this theory of the behaviour of the

word 'true' can derive everything we need to say about truth. In this section I argue that when deflationism is understood in these terms, all axiomatic theories of truth we have are deflationary theories of truth. This is a strong claim, but one which I believe follows from my understanding of deflationism.

If we look at common axiomatic theories of truth that have been provided (TB, CT, FS, KF, etc.), we see that these theories only depend upon logical, linguistic and semantic notions. They make use of common logical constants (such as the connectives and quantifiers), common logical inference rules and introduce a new syntactic symbol Tr. They rely upon arithmetisation of syntax and other such syntactic operations and the linguistic/semantic notions of what a sentence or formula is. These theories rely on nothing further than these LLS notions. No axiomatic theory of truth makes reference to a new class of entities such as 'facts', substantive metaphysical connectives such as 'corresponds', nor even significant second order resources such as a comprehension scheme to quantify over properties. An axiomatic theory of truth is solely given by axioms in the language of the background syntax and the new symbol for the truth predicate. If anything should count as an LLS theory of truth, an axiomatic theory of truth is a prime example of such a theory. Axiomatic theories of truth make no reference to other substantive philosophical notions, other than those logical ones already provided by the background theory.

I argue that thus we should regard all of these axiomatic theories as deflationary theories of truth. It is hard to see what is non-deflationary about these theories. At no point are 'substantive' notions made use of, and at no point do we see the statement of the theory require resources significantly beyond the background logical ones. Axiomatic theories of truth are as insubstantial a theory of truth as one can offer, whilst retaining the truth predicate's role of semantic ascent and descent. Logical resources are required, as are certain syntactic abilities, but an axiomatic theory requires no further resources to be formulated beyond these.

In the more philosophical theories of truth, we see evidence for this claim. Common deflationary theories of truth are presented in a similar axiomatic way. For example, Horwich's (1998) minimalist theory of truth is very similar to the axiomatic theory of TB, but propositions are the bearers of truth and the theory contains additional 'non-controversial' instances of the equivalence schema for propositions also containing the truth predicate. As stated in Section 4.4, Quine's (1986) theory of truth is very similar to Tarski's compositional theory of truth CT.

These are offered as deflationary theories of truth, in part, for their logical axiomatic structure. If the truth predicate for our base language B can be given solely by these axioms in the language $\mathscr{L}_B \cup \{Tr\}$, then the notions this predicate relies upon offer no substance beyond $\mathscr{L}_B$ and an additional syntactic symbol. This certainly doesn't make truth a 'substantive', 'weighty' notion, even if it allows one to prove new consequences in their theory, or is not invariant under certain similarity relations. An axiomatic theory of truth is by formulation a deflationary theory of truth. A theory of truth formed by axioms is formed only of LLS notions.

Some clarification should be offered here. I do not wish to say that any axiomatic theory is a deflationary theory. This would be a hasty generalisation. I restrict these remarks only to theories of truth, and whether axiomatic theories are deflationary in general I leave open. Considering the axiomatic theory of set theory ZFC as an example, it is hard to know what it would even mean for a set theory to be deflationary. As acknowledged in Chapter 3, deflationism is a term of art and its use for mathematical theories is perhaps unfixed. We could understand deflationist set theory to mean the same as I argue it does for truth - a logical-linguist-semantic theory of 'set'. In this case, perhaps ZFC is a deflationary theory of sets, if the minimal notions it relies upon are logical-linguistic-semantic in nature. This is in contrast to a different theory of collections, such as van Inwagen's (1990) organicist theory of mereology – that xs compose an object y if and only if the activity of the xs constitutes a life. This theory is not deflationary, since it relies upon an understanding of activity and life, neither of which appear to be logical-linguistic-semantic concepts. Such discussions go well beyond my current claim, however, and I only claim that axiomatic theories of truth are deflationary, in the sense of deflationist as it applies to truth.

It would also be too hasty to conclude that the converse holds and all deflationary theories of truth are axiomatic theories of truth. It appears that logical-linguistic-semantic is a broad enough notion that we could have theories of truth which depend solely upon LLS notions, and yet are not axiomatic theories of truth. Some semantic theories of truth could be an example here, as well as broader philosophical theories of truth not expressible axiomatically. I leave the possibility of this open as well, for this chapter is exploring the connection between deflationary theories of truth and formal theories of truth only.

A further important clarification, is that an axiomatic theory of truth ought to be a theory *of truth*, and not just a theory involving a truth predicate. McGee's

trick allows one to use the diagonal lemma to construct a sentence $\sigma \leftrightarrow (\gamma \leftrightarrow Tr(\ulcorner\sigma\urcorner))$ which is equivalent to $\gamma \leftrightarrow (Tr(\ulcorner\sigma\urcorner \leftrightarrow \sigma)$, and we can do this for any γ in $\mathscr{L}_{Tr}$ (Raatikainen, 2006). This allows one to construct a theory PA$+\{Tr(\ulcorner\sigma_i\urcorner) \leftrightarrow \sigma_i : i \in \mathbb{N}\}$ which has the same arithmetical consequences as any arbitrarily strong axiomatisable theory extending PA. I do not wish to claim that the arithmetical consequences of this theory is a deflationary theory of truth, for clearly these McGeean-Biconditionals $Tr(\ulcorner\sigma_i\urcorner) \leftrightarrow \sigma_i$ do not constitute a theory of truth. A theory of truth ought to account for the ways in which we appear to use the truth predicate linguistically, and these biconditionals do no such thing. What exactly constitutes a theory of truth is not an easy matter to ascertain,[15] but there are certainly some examples which seem to certainly be truth-like theories (e.g. those axiomatic theories mentioned above) and others involving a truth-like predicate which are not theories of truth (such as one formulated of McGeean-Biconditionals).

The other important point to make is that just because we have axiomatic theories of truth, and that these should be regarded as deflationary theories of truth, does not automatically support deflationism about truth. There remains the wider question of whether any of these theories of truth are actually good enough as theories of truth. If it can be shown that no axiomatic theory of truth is adequate as a theory of truth, then this is a significant blow to deflationism about truth. If, on the other hand, an axiomatic theory of truth is offered that provides rules of use for the truth predicate which are liberal enough for all the deflationists' needs, then it appears that we do have an adequate theory of truth. I thus see that the success of deflationism about truth rises or falls with the success of axiomatic theories of truth. In particular, an adequate theory of the latter kind suggests an adequate theory of the former kind.

This leads to the debate into a new area - of whether we actually have a suitable axiomatic theory of truth, and just what properties we require a theory of truth to have. Do we want a theory of truth to be classical, particularly if the background logic is, or is this not so important? Should a theory of truth derive all the T-schema, or is there some restriction upon this? Should a theory of truth satisfy generalised compositional principles, or is this restricted also? If we are

[15]Leitgeb (2007) provides a set of jointly unsatisfiable norms for a theory of truth and Horsten and Halbach (2015) provide alternative desiderata for a theory of truth. So far what exactly we want from a theory of truth formally appears to be an under-explored area ready for further research.

to argue in the affirmative for all of these, then it seems like deflationism about truth is incorrect, for our closest theory to satisfying all of these properties, FS, is ω-inconsistent, and thus has no model over the natural numbers. If the choice is between deflationism about truth or the standard model of arithmetic, it appears like an easy choice. If there is some leeway with these properties, then perhaps an axiomatic theory can be an adequate theory of truth, providing great support for a deflationary position about truth. I leave this then debate open for now, for I would like to conclude with some remark upon semantic theories of truth.

My proposal is that we should regard all axiomatic theories of truth as deflationary theories of truth, but this is only one half of the study of formal theories of truth, and ignores semantic theories of truth - such as satisfaction classes and Kripkean fixed points. Can a semantic theory of truth be a deflationary theory of truth?

I suggest that firstly we should be clear on our terminology here. A satisfaction class or a Kripkean fixed point is not a theory of truth, but instead an interpretation of a truth predicate, or alternatively a class of 'true' sentences. One particular interpretation seems to suggest little immediately about deflationism or not, when deflationism is portrayed as I do in Chapter 3 as interested in the behaviour of the word 'true'. It appears possible for two different theories of truth (a minimalist theory and a correspondence theory, say) to believe that the same sentences are true (ignoring sentences about the theories themselves), but have radically different views on what the word 'true' means. It would not be correct to class an interpretation alone as deflationary, or not deflationary, for it is the theory of that interpretation that is deflationary or not.

What we might be able to judge as deflationary or not is a general semantic process for constructing particular interpretations of the truth predicate. For instance, the model-theoretic method of assembling satisfaction classes via end extensions, or of building fixed points via Kripke-jump operators. Whether these methods are deflationary or not I think depends too much upon the construction to provide a general answer. Each construction would have to be evaluated on its own merits and tested against the limits of what we regard as logical-linguistic-semantic.[16] I

[16]Soames (1998, Ch. 8) argues that both Tarski and Kripke's semantic theories are deflationary theories, since their general rules ensure truth is not a "contentious metaphysical or epistemological notion". It seems that Soames regards the formal machinery behind these theories as in essence just formal explications of linguistic rules. I have no critique of this analysis, but am hesitant to conclude that set-theoretic notions essential to the theories are 'just' formal

expect that in some cases, the answer would be that they are deflationary, and in other cases, the answer would be negative. I anticipate that only hard philosophical analysis, and not a quick criterion of conservativity or otherwise, would be capable or providing such a judgement.

In conclusion, we ought not regard either proof-theoretic conservativity, model-theoretic conservativity or logicality as good tests for deflationism. For the case of semantic theories of truth, it appears like there is no easy answer of whether these theories are deflationary or not, and only hard conceptual analysis of each theory will provide the answer. On a more positive note, however, I argue that one class of formal theories of truth should be regarded as deflationary. These are the axiomatic theories of truth, for they depend solely upon logical-linguistic-semantic notions. The debate around deflationism should not focus on what formal properties these theories have as a test of deflationism (as interesting as these results are), but instead whether these theories are adequate as theories of truth at all. If we have an adequate axiomatic theory of truth, then we have an adequate deflationary theory of truth. This leads to the next major question of this thesis: is it possible to have an axiomatic theory of truth that is adequate?

The next two chapters of this thesis are inspired by this question as to whether any axiomatic theory of truth is adequate as a theory of truth. I here distinguish between two types of adequacy: formal adequacy and philosophical adequacy. Formal adequacy is a theory which is formally consistent, captures as many syntactic features of the truth predicate as possible, such as an extended T-Schema from Chapter 2, and has a suitable model-theoretic interpretation. Philosophical adequacy is a theory of truth which can provide for, or explain away, common philosophical uses of truth. In Chapter 5 I will introduce and explore two new axiomatic theories of truth, with the aim to show that we have at least one formally adequate axiomatic theory of truth. This chapter will explore these theories' formal features, but also their philosophical implications. I will conclude that we can regard the first of these as an adequate formal theory of truth, providing support for deflationism about truth, and partially answering the question above.

explications, and thus do not carry important content on their own.

Chapter 5

Axioms for Truth: Two Novel Theories of Truth and Paradox

In the preceding chapter I argued that we should regard all our current axiomatic theories of truth as deflationary theories of truth. This means that to research the primary question of this thesis further, whether formal theories of truth support deflationism or not, it needs to be examined whether there is an adequate axiomatic theory of truth. At the end of Chapter 4 I distinguished between two types of adequacy: formal and philosophical. This chapter introduces two new axiomatic theories of truth and will conclude that the first of these is adequate in the formal sense. Here I take formal adequacy to mean that the theory can provide the truth predicate's syntactic features (including an extended T-Schema from Chapter 2, avoids semantic paradoxes and has at least one suitable interpretation. For the most part this chapter shall leave this motivation in the background, leaving space to explore the formal details of these theories uninterrupted, but these results will inform the adequacy of the first theory of truth. This suggests formal theories of truth support deflationism and provides a partial answer to the question at the end of Chapter 4, with the remainder, and further support, to be provided by Chapter 6.

Chapter Abstract

In this chapter I present two new axiomatic theories of truth. The first is an axiomatic typed theory of truth, denoted ATT, which can quantify

over the levels of the typed truth hierarchy. The theory contains a single binary truth predicate which has the interpretation that a sentence is true at a particular level of the hierarchy. This theory enables proof of sentences about the whole hierarchy within the theory itself, and also internally proves that various Liar-like sentences are not true. I provide a philosophical interpretation of this theory and relate it to debates in philosophical logic about the semantic paradoxes and absolute generality. The second theory of truth I present is a type-free KF-like theory of truth which is analogous to ATT as KF is to $\mathrm{RT}_{<\epsilon_0}$. I show how a new three-valued logic motivates alternative axioms to KF and provide the details of this theory. I prove numerous results about this theory, discuss its relation to KF, and remark upon the relation this shows between typed and type-free theories of truth.

5.1 Introduction

In this chapter I present two new axiomatic theories of truth. In Section 5.2 I introduce the first theory of truth: axiomatic typed truth (ATT). This is a typed theory of truth in the Tarskian tradition, where the truth of a sentence is evaluated only relative to a particular rank in a hierarchy. This theory contains only a single binary truth predicate, however, which tracks both sentences and the level of evaluation for the sentences. Each sentence is assigned a rank and all our naive classical alethic reasoning holds for those sentences whose rank is a strictly positive ordinal. I argue that sentences with rank 0 should be interpreted as not 'truth-apt' and these fall into two main categories: those which quantify over absolutely all levels of the truth predicate, and those which are paradoxical.

In Section 5.3 I present the second theory of truth (KFJ). This theory is similar to KF, but has novel axioms for the truth of a negated conjunctive sentence and a negated disjunctive sentence. I show that this theory is motivated by a new three-valued logic ML_3, in the same way that KF is motivated by the Strong Kleene scheme. This theory is type-free, and can be seen as something of a 'type-free' variant of ATT, analogous to the relationship between KF and ramified truth up to ϵ_0 ($\mathrm{RT}_{<\epsilon_0}$). In particular, in Section 5.4 I show that if ATT proves a formula is true, then KFJ will also prove that this formula is true. The converse does not hold, however, as KFJ loses a distinction between truth-apt and non-truth-apt sentences. The benefit of this is that KFJ views sentences which quantify over

absolutely all truth predicates as true or false, but the drawback is that KFJ no longer provides a resolution of the semantic paradoxes.

Throughout this chapter I will treat Gödel codes of sentences as the bearers of truth, although often I will simply write a sentence σ rather than its Gödel code $\ulcorner\sigma\urcorner$ for readability. I will also make use of the dot notation outlined in Section 1.2, which allows for easy presentation of codes of formulas. Context should make it clear when I refer to a sentence, and when I am referring to the code of a sentence, even if the notation is not always explicit. For the details of this notation system, as well as Gödel codes more generally, the reader is referred to Halbach's *Axiomatic Theories of Truth* (Halbach, 2011, p. 32-33). Finally, I note that these theories are presented for a language $\mathscr{L} \supseteq \mathscr{L}_A$ and $B \supseteq$ PA. This is certainly more strength than required, but ensures that we have adequate resources for all the syntactic operations required of the coding.

5.2 Axiomatic Typed Truth

5.2.1 Motivation

The Liar paradox poses a notorious problem for any axiomatic theory of truth: our naive norms for a theory of truth cannot be mutually satisfiable. Typically one wishes to have a theory of truth which satisfies a full T-Schema ($\ulcorner\varphi\urcorner$ is true iff φ) and behaves only according to classical logic – both internally and externally. It is well-known that adding such a theory of truth to only a minimal theory of syntax results in a Liar paradox and hence inconsistency.

Tarski's solution to this problem is to introduce type-restrictions to the truth predicate (Tarski, 1956). Rather than introducing a single truth predicate to the language, which is predicated of all sentences in the language (including itself), we have a family of truth predicates, ordered hierarchically. If we wish to say that a sentence containing a particular truth predicate is true or not true, we express this with a truth predicate which is strictly higher in the hierarchy. In this way, formal consistency is achieved, but at the cost of certain philosophical desiderata – particularly the desideratum that we have only a single truth predicate in our language which suffers from no type restrictions.

One interpretation of typed theories of truth is that there is no notion of 'absolute' truth, but instead all truth is relative to a certain rank. Glanzberg

(2001) provides philosophical justification of this using the notion of 'context shift' from philosophy of language . Contextual theories of truth take truth to be a contextual notion: its behaviour can shift according to the linguistic context in which it is uttered. Under such theories a sentence can be true in context A (e.g. it is true that the sun will rise tomorrow in the context of planning what to do for the weekend) but not true in another context B (e.g. the same sentence may not be true in a sceptical philosophy seminar). Often these theories have been formally modelled by Tarskian hierarchical theories of truth with a plurality of truth predicates, such as by Burge (1979).

One key issue with the formal Tarskian approach is that in natural language we appear to only have a single truth predicate. Tarski's approach requires an introduction of (often) transfinitely-many truth predicates to the language, all of which are distinct. These truth predicates are then organised by meta-level variables, which cannot be discussed in the object language directly. This means that we cannot prove certain natural properties of a typed theory of truth in the object language itself, for instance: for any successor ordinal i, if a sentence σ is true according to truth predicate labelled $i-1$, then it is true according to the truth predicate labelled i. This cannot be proven inside the object language, since we cannot quantify over such i's, but is provable in the metalanguage.

In this section I will introduce a typed theory of truth which overcomes these limitations. I call this theory Axiomatic Typed Truth, which is abbreviated ATT. I introduce a typed theory of truth, similar to Tarski's theory of truth, but where the typing occurs as an object-level variable inside the language. The theory introduces only a single truth predicate to the language, but this is a binary predicate of both a sentence and a level of evaluation. Because the types of the truth predicate are object-level variables, these can be quantified over in the theory itself, and the theory can prove statements about these. This produces a consistent theory of truth which retains highly classical behaviour, and much of the T-Schema, whilst avoiding the Liar paradox. I hope that this theory at the very least offers a better formal approximation of the contextualist approach to truth, given these advantages.

In presenting this theory we work with a language $\mathscr{L} \supseteq \mathscr{L}_A$ and a theory B which is sufficiently strong enough to interpret PA. This ensures that it has the required strength to perform the necessary coding, syntactic, and ordinal operations. We extend $\mathscr{L}$ to $\mathscr{L}_{Tr} = \mathscr{L} \cup \{Tr(x,y), R(x,y)\}$. The relation $Tr(x,y)$ is

intended as our truth predicate for the language, where x ranges over Gödel codes of sentences and y ranges over codes of ordinals, which track the level of truth evaluation in the hierarchy. The new relation $R(x, y)$ is intended as a notion of 'rank' for sentences. This will be introduced and discussed in the following section.

5.2.2 Rank

I introduce ATT by presenting the notion of the rank of a formula. The rank of a formula defines which level of the hierarchy a sentence should be evaluated at, and informally the relation R does this by tracking the highest level of truth predicates contained in the formula. These levels are denoted by codes of ordinals, which allows our hierarchy to extend into the transfinite. This ordinal coding covers ordinals up to, but not including, ϵ_0, since PA can prove the well-ordering of these ordinals. Sentences of the base language $\mathscr{L}$ are given rank $\mathring{1}$ and sentences using a truth predicate at rank $\mathring{n}$ are given rank $(n\mathring{+}1)$. Formulas which are deemed as 'pathological' are given rank 0 and these will be discussed in Section 5.2.5.

Our expanded language $\mathscr{L}_{Tr}$ contains a relation $R(x, y)$ which pairs every (code of an) $\mathscr{L}_{Tr}\backslash\{R\}$ formula x with a unique (code of an) ordinal y. The intended interpretation of this relation is that every truth-apt formula will be assigned a non-zero ordinal which is the level its truth or falsity should be evaluated at. I will be somewhat informal with ordinal notation, and often simply write the ordinal n itself, rather than what should strictly be the numeral which codes that ordinal, which I denote $\mathring{n}$. We note that our base theory B is sufficiently rich to code ordinal operations $\mathring{+}$, $\mathring{<}$, $\mathring{>}$ and $\mathring{=}$ as well as $max\{a, b\}$ and $sup\{a_i : i \in I\}$ which return the maximum of two ordinals a and b and the supremum of a family of ordinals a_i respectively. We explicitly note that if we cannot provably code the supremum of a collection of ordinals, then the result of the supremum function on this collection is $\mathring{0}$. We also note that B can provably define the predicates $At_{\mathscr{L}}(x)$ and $Form(x)$, which express "x is the code of an atomic $\mathscr{L}$-formula" and "x is the code of a formula" respectively.

The formula $R(x, y)$ is defined by the following axioms:

(R1): $\forall x[At_{\mathscr{L}}(x) \rightarrow R(x, \mathring{1})]$

(R2): $\forall x[Form(\dot{\neg}x) \rightarrow (R(x, y) \leftrightarrow R(\dot{\neg}x, y))]$

(R3): $\forall a \forall b[Form(a\dot{\wedge}b) \rightarrow (R(a\dot{\wedge}b, y) \leftrightarrow y = \max\{n, m : R(a, n) \wedge R(b, m)\})]$

(R4): $\forall a \forall b[Form(a\dot{\vee}b) \rightarrow (R(a\dot{\vee}b, y) \leftrightarrow y = max\{n, m : R(a, n) \wedge R(b, m)\})]$

(R5): $\forall x[Form(\dot{\forall}ax(a)) \rightarrow (R(\dot{\forall}ax(a), y) \leftrightarrow y = sup\{y_i : R(x(i), y_i)\})]$

(R6): $\forall x[Form(\dot{\exists}ax(a)) \rightarrow (R(\dot{\exists}ax(a), y) \leftrightarrow y = sup\{y_i : R(x(i), y_i)\})]$

(R7): $\forall x[Form(\dot{Tr}(x, n)) \rightarrow (R(\dot{Tr}(x, n), y) \leftrightarrow$
$$\leftrightarrow y = \begin{cases} max\{n+1, z+1\} & : R(x, z) \wedge z > 0 \\ 0 & : Otherwise \end{cases}\Big)\Big]$$

(R8): $\forall x \forall y[R(x, y) \rightarrow \forall n \mathring{\neq} y(\neg R(x, n))]$

(R9): Else, $R(y, \mathring{0})$.

We can think of the rank of a formula as tracking the levels of truth predicates in the formula, where a rank of $n > 0$ denotes that $n - 1$ is the highest level in the truth hierarchy utilised by the formula. If a formula is assigned a non-zero rank n, then it behaves 'nicely'. Either that formula or its negation will be determined to be true at level $n + 1$ in the hierarchy, and classical equivalences will hold for this formula. A rank of 0 is only given when this procedure for assigning a formula a rank will not terminate, which shows that the formula is not well-founded in some sense. A canonical example of a formula with Rank 0 is the liar paradox, but we shall see other examples in Section 5.2.5.

The following lemma shows that all sentences of our base language $\mathscr{L}$ are given rank 1, and hence the only formulae of higher (or lower) rank involve the truth predicate in some way.

Lemma 5.2.2.1. *σ is an $\mathscr{L}$-formula if and only if $R(\sigma, \mathring{1})$.*

Proof. We prove this via induction on the complexity of σ. We know σ is an atomic $\mathscr{L}$-formula if and only if $R(\sigma, \mathring{1})$ by R1.

- Suppose σ is of the form $\neg\eta$. We know by induction that $R(\eta, \mathring{1})$ and hence by R2 $R(\sigma, \mathring{1})$.
- Suppose σ is of the form $\alpha \wedge \beta$. We know by induction that $R(\alpha, \mathring{1})$ and $R(\beta, \mathring{1})$ and hence by R3 $R(\sigma, \mathring{1})$.
- Suppose σ is of the form $\alpha \vee \beta$. We know by induction that $R(\alpha, \mathring{1})$ and $R(\beta, \mathring{1})$ and hence by R4 $R(\sigma, \mathring{1})$.

- Suppose σ is of the form $\exists x \eta(x)$. We then know by induction that $R(\eta(i), \mathring{1})$ for each i and hence by R5 $R(\sigma, \mathring{1})$.
- Suppose σ is of the form $\forall x \eta(x)$. We then know by induction that $R(\eta(i), \mathring{1})$ for each i and hence by R5 $R(\sigma, \mathring{1})$.

□

We will see later, in Lemma 5.2.4.5, that a useful feature of the definition of rank is that if a sentence is given a rank $a > 1$, then it is logically equivalent to a sentence also of rank a of the form $Tr(x, a-1)$ where x is a formula with rank $a-1$. This tells us that, in a sense, we can 'pull' a truth predicate to the outside of the formula, without affecting its logical or alethic status. This requires axioms governing the truth predicate to prove, and thus this result is delayed until Section 5.2.4.

This definition of rank becomes the formal framework from which we build our theory of truth, ATT. We leave the notion of rank itself alone now, but it is an interesting open question to consider what it's recursion-theoretic complexity is.

Question 5.2.2.2. *What is the recursive complexity of the rank relation R?*

5.2.3 Axioms of ATT

With the definition of rank in place we can now define the truth predicate of our theory of axiomatic typed truth. In the language $\mathscr{L}_{Tr}$ our truth predicate is $Tr(x, y)$ which has the intended interpretation that (the Gödel code of) a sentence x is true at (the ordinal code of) a level y. The level y is the typing of the truth predicate, and a formula cannot be true at level y unless it has rank less than or equal to y.

Let Tr_{At} be the partial (definable) truth predicate for $\mathscr{L}$-atomic sentences. We note that B is sufficiently strong to provably code the predicates $Sent(\sigma)$ and $Ord(n)$ which express "σ is the Gödel code of a ($\mathscr{L}_{Tr}$) sentence" and "n is the code of an ordinal" respectively. The Axiomatic Typed Truth (ATT) predicate $Tr(\sigma, y)$ is given by the following axioms:

(ATT1): $\forall \sigma [At_{\mathscr{L}}(\sigma) \rightarrow (Tr(\sigma, \mathring{1}) \leftrightarrow Tr_{At}(\sigma))]$.

(ATT2): $\forall \alpha \forall \beta \forall n [Sent(\alpha \dot{\wedge} \beta) \rightarrow (Tr(\alpha \dot{\wedge} \beta, n) \leftrightarrow Tr(\alpha, n) \wedge Tr(\beta, n))]$

(ATT3): $\forall\alpha\forall\beta\forall n[Sent(\alpha\dot{\vee}\beta) \rightarrow (Tr(\alpha\dot{\vee}\beta, n) \leftrightarrow Tr(\alpha, n) \vee Tr(\beta, n))]$

(ATT4): $\forall\sigma\forall n[(Sent(\dot{\neg}\sigma) \wedge R(\sigma, n) \wedge n\mathring{>}\mathring{0}) \rightarrow (Tr(\dot{\neg}\sigma, n) \leftrightarrow \neg Tr(\sigma, n))]$

(ATT5): $\forall\sigma\forall n[Sent(\dot{\forall}x\sigma(x)) \rightarrow (Tr(\dot{\forall}x\sigma(x), n) \leftrightarrow \forall a\, Tr(\sigma(\dot{a}), n))]$

(ATT6): $\forall\sigma\forall n[Sent(\dot{\exists}x\sigma(x)) \rightarrow (Tr(\dot{\exists}x\sigma(x), n) \leftrightarrow \exists a\, Tr(\sigma(\dot{a}), n))]$

(ATT7): $\forall\sigma\forall n[Sent(\dot{Tr}(\sigma, n) \rightarrow (Tr(\dot{Tr}(\sigma, n), (n\mathring{+}1)) \leftrightarrow Tr(\sigma, n))]$

(ATT8): $\forall\sigma\forall n[Tr(\sigma, n) \rightarrow \forall k[(Ord(k) \wedge k\mathring{>}n) \rightarrow Tr(\sigma, k)]]$

(ATT9): $\forall\sigma\forall y[R(\sigma, y) \rightarrow \forall k[(Ord(k) \wedge k\mathring{<}y) \rightarrow \neg Tr(\sigma, k)]]$

(ATT10): $\forall\sigma[R(\sigma, \mathring{0}) \rightarrow \forall y\neg Tr(\sigma, y)]$

(ATT11): $\forall\sigma\forall n[Tr(\sigma, n) \rightarrow (Sent(\sigma) \wedge Ord(n))]$

(ATT12): An induction axiom for every formula in the language $\mathscr{L}_{Tr}$.

The axioms provide a typed theory of truth, where axiom ATT1 defines truth for atomic formulas and axioms ATT2-6 provide near-classical compositionality. ATT7 is our axiom for speaking of the truth of sentences which themselves contain a truth predicate and crucially ensures that a sentence talking about truth at rank n is evaluated at rank $n + 1$. Axioms ATT8-11 detail important structural properties of the truth predicate. ATT8 ensures that that the truth predicate behaves monotonically (if a sentence is determined to be true at level n, then it will remain true at higher levels). The next two axioms detail how the rank relation is crucial in understanding the truth predicate. ATT9 tells us that a sentence will never be true at a level lower than its rank and ATT10 details that rank 0 sentences will never be true at any level. ATT11 provides the structural property that only Gödel codes of sentences are true, and these are only true at levels which are codes of ordinals. Finally ATT12 provides an induction scheme for $\mathscr{L}_{Tr}$, which enables proofs of many desired properties for the theory in general.

We note that the compositional axiom for negation ATT4 has the important caveat that a negated sentence is true only if it has a non-zero rank. Were this not to be the case the theory would be inconsistent. This is due to ATT9 which expresses that all sentences of rank 0 are not true and since the negation of a rank 0 sentence also has rank 0 both a rank 0 and its negation are not true. This

restriction means that the wholly general compositional principle $Tr(\alpha\dot{\rightarrow}\beta, n) \leftrightarrow (Tr(\alpha, n) \rightarrow Tr(\beta, n))$ does not hold.[1] Consider a formula α such that $R(\alpha, n+1)$, then we have $R(\alpha\dot{\rightarrow}\beta, n+1)$ as well by R2 and R3, and hence $\neg Tr(\alpha\dot{\rightarrow}\beta, n)$ by ATT9. We do have $Tr(\alpha, n) \rightarrow Tr(\beta, n)$ by pure logic, however, since $\neg Tr(\alpha, n)$. The notion of rank tells us that we cannot evaluate the truth of a conditional sentence at rank n in this instance, and it requires evaluation at level $n+1$. We thus have to restrict compositionality for the material conditional to evaluation at the level of the antecedent, as the following lemma shows.

Lemma 5.2.3.1.

(ATT13): $\forall\alpha, \beta, n[R(\alpha, n) \rightarrow (Tr(\alpha\dot{\rightarrow}\beta, n) \leftrightarrow (Tr(\alpha, n) \rightarrow Tr(\beta, n)))]$

Proof. We work inside B + ATT, assume that (codes of) formulas α and β are given and suppose $R(\alpha, n)$. We know that $Tr(\alpha\dot{\rightarrow}\beta, n) \leftrightarrow Tr(\neg\alpha, n) \vee Tr(\beta, n)$ by ATT3. Since $R(\alpha, n)$ we thus have $Tr(\alpha\dot{\rightarrow}\beta, n) \leftrightarrow \neg Tr(\alpha, n) \vee Tr(\beta, n)$ by ATT2. □

There is a similar restriction on compositionality for the biconditional. To evaluate whether a biconditional is true at level n both immediate-subformulae require rank at least n. The compositional rule for the biconditional is stated below, although the proof is routine and hence omitted.

Lemma 5.2.3.2.

(ATT14): $\forall\alpha, \beta, n[(R(\alpha, n) \wedge R(\beta, n)) \rightarrow (Tr(\alpha\dot{\leftrightarrow}\beta, n) \leftrightarrow (Tr(\alpha, n) \leftrightarrow Tr(\beta, n)))]$

These two lemmas can be seen as derived rules for the behaviour of the material conditional and biconditional with the truth predicate, and are hence designated as ATT13 and ATT14 respectively.

We now show that the theory ATT is consistent over first order Peano Arithmetic (PA) by building an appropriate model construction, before considering further features of the theory. This model is Tarski's semantic typed theory of truth with truth predicates up to ϵ_0, and an appropriate interpretation of the rank relation.

[1] We explicitly assume that $\alpha \rightarrow \beta$ is a shorthand for the formula $\neg\alpha \vee \beta$, and that $\alpha \leftrightarrow \beta$ is a shorthand for $(\alpha \rightarrow \beta) \wedge (\beta \rightarrow \alpha)$.

Theorem 5.2.3.3. PA + ATT *is consistent, and moreover, Tarski's semantic theory of typed truth predicates up to* ϵ_0*, along with an appropriate rank function, over* $\mathbb{N}$*, is a model of* PA + ATT*.*

Proof. We take the standard natural numbers $\mathbb{N} \vDash$ PA and work in the language $\mathscr{L}_A \cup \{T^i : i < \epsilon_0\}$, where each T^i is interpreted as a Tarskian truth predicate. Explicitly, take $T^1 \vDash \mathrm{CT}(\mathbb{N})$ and $T^{n+1} \vDash \mathrm{CT}((\mathbb{N}, T^1, \ldots, T^n)$ for all successor ordinals n. At a limit ordinal λ we define $T^\lambda = \bigcup_{i<\lambda} T^i$.

First we build an interpretation of the rank function $R(\sigma, n) = \{(\sigma, n)\} \subseteq \mathbb{N}\times\mathbb{N}$ as a set of pairs by an induction on (codes of ordinals) n. If σ codes an $\mathscr{L}_A$-formula, then $(\sigma, \mathring{1}) \in R$. If σ codes an $\mathscr{L}_A \cup \{T^n\}$-formula, where n is 1 or a successor ordinal, then $(\sigma, (n \mathring{+} 1)) \in R$. Finally, if σ codes an $\mathscr{L}_A \cup \{T^\lambda\}$-formula, where λ is a limit ordinal, then $(\sigma, \mathring{\lambda}) \in R$. We continue this construction up to ϵ_0, and then for each $a \in \mathbb{N}$ with no y such that $(a, y) \in R$ we have that $(a, \mathring{0}) \in R$.

With this interpretation of $R(\sigma, n)$ we are able to provide our model of PA + ATT with the interpretation $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i)$, where $Tr(\sigma, n)$ is interpreted as $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\sigma)$ for each n. We identify quantified $\mathscr{L}_{Tr}$-sentences of the form $\exists x\, Tr(\sigma, x)$ and $\forall x\, Tr(\sigma, x)$ with their metalanguage equivalents, i.e. $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^x(\sigma)$ for some x and $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^x(\sigma)$ for every x.

We now show that this model satisfies the axioms of the truth predicate for ATT. We do not show that the rank axioms are satisfied, since this is implicit in our construction of R.

(T1): If σ is (the Gödel code of) an atomic formula in $\mathscr{L}_A$, then we have that $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^1(\sigma) \leftrightarrow Tr_{At}(\sigma)$

(T2): $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\alpha \dot{\wedge} \beta)$ if and only if $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\alpha) \wedge T^n(\beta)$.

(T3): $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\alpha \dot{\vee} \beta)$ if and only if $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\alpha) \vee T^n(\beta)$.

(T4): Suppose that $(\sigma, n) \in R$ and $n \mathring{>} \mathring{0}$. Hence σ codes a formula from $\mathscr{L}_A \cup \{T^1, \ldots, T^{n-1}\}$. We have that $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\sigma) \leftrightarrow \neg T^n(\neg\sigma)$.

(T5): $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\forall x \varphi(x)) \leftrightarrow \forall a T^n(\varphi(\dot{a}))$.

(T6): $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\exists x \varphi(x)) \leftrightarrow \exists a T^n(\varphi(\dot{a}))$.

(T7): $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\sigma) \leftrightarrow T^{n+1}(\ulcorner T^n(\sigma) \urcorner)$.

(T8): If $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\sigma)$ then, for each $i > n$, $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^i(\sigma)$

(T9): Suppose $(\sigma, y) \in R$. If $y \dot{=} \mathring{0}$ the axiom holds vacuously, so we assume $y \dot{>} \mathring{0}$. Therefore σ codes an $\mathscr{L}_A \cup \{T^1, \ldots, T^{y-1}\}$-formula and σ is not an $\mathscr{L}_A \cup \{T^1, \ldots, T^{y-2}\}$-formula. Therefore, due to Tarski's typed construction we have that $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash \neg T^k(\sigma)$ for each $k < y$.

(T10): If $(\sigma, \mathring{0}) \in R$, then σ does not code a formula from $\mathscr{L}_A \cup \{T^i : i < \epsilon_0\}$ and therefore $\bigcup_{i<\epsilon_0}(\mathbb{N}, T^i) \vDash \neg T^k(\sigma)$ for every k, since each truth predicate is a property of only sentences.

(T11): By the details of our construction we know that if $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i) \vDash T^n(\sigma)$, then n is an ordinal and σ is the Gödel code of a sentence.

(T12): Since $\bigcup_{i<\epsilon_0}(\mathbb{N}, R, T^i)$ is an expansion of the natural numbers, we know that induction holds for this model as well.

We thus have that $\bigcup_{n\in\mathbb{N}}(\mathbb{N}, R, Tr^n) \vDash$ PA + ATT and hence PA + ATT is consistent. □

We hence see that there is great similarity between the theories of PA + ATT and PA + $\mathrm{RT}_{<\epsilon_0}$, and this will be detailed and proven later in Theorem 5.2.4.6.

The interesting aspect of ATT, and where it differs to $\mathrm{RT}_{<\epsilon_0}$, is that the rank variable in the truth predicate is a variable in the object language, rather than a metalevel variable. This allows quantification over levels of the truth predicate and enables theorems about the truth hierarchy to be stated and proven in the theory itself. This is desirable for those who wish for a treatment and discussion of truth to be given without needing to rely on external metalanguage resources - in particular this seems closer to the way we use the truth predicate in natural language, where we have no metalevel resources to draw upon. In the following section I will present and prove a number of key theorems about the behaviour of ATT and the desirable and sometimes novel behaviour that results from treating levels of the truth predicate as object-language variables.

5.2.4 Alethic Features of ATT

One of ATT's main strengths as a typed theory of truth is its ability to quantify over levels of the truth predicate, and this results in a number of interesting features. I shall state and prove many of these here. Axiom ATT8 tells us that truth

is monotone increasing, in the sense that if a sentence is true at a level n, then it is true at every level higher than this. We are able to prove (within the theory itself) that truth is somewhat monotone decreasing as well, and if a sentence is true at a level n, then it is true at each level below this, as long as the level is not lower than that sentence's rank.

Lemma 5.2.4.1. $\mathrm{B} + \mathrm{ATT} \vdash \forall\sigma\forall n[(\mathrm{Tr}(\sigma, n) \wedge R(\sigma, a)) \rightarrow$
$\rightarrow \forall k[(Ord(k) \wedge (a \mathring{\leqslant} k \mathring{\leqslant} n)) \rightarrow \mathrm{Tr}(\sigma, k)]]$

Proof. Suppose this does not hold, then there is a k where $a \leqslant k \leqslant n$ such that $\neg Tr(\sigma, k)$, but $Tr(\sigma, n)$ and $R(\sigma, a)$. By ATT4 we deduce $Tr(\neg\sigma, k)$ and hence by ATT8 $Tr(\neg\sigma, n)$. We again use ATT4 to deduce $\neg Tr(\sigma, n)$ which is a contradiction. □

One of the useful alethic features of axiomatic typed truth is that it provides a T-Schema for every sentence with rank greater than 0. This provides Tarskian material adequacy for all sentences the theory views as reasonable, and enables the proof of many desirable features of a theory of truth.

Theorem 5.2.4.2 (T-Schema)**.** *For any $\mathscr{L}_{\mathrm{Tr}}$ formula σ:*
$\mathrm{B} + \mathrm{ATT} \vdash \forall n[(R(\ulcorner\sigma\urcorner, n) \wedge n \mathring{>} \mathring{0}) \rightarrow (\sigma \leftrightarrow \mathrm{Tr}(\ulcorner\sigma\urcorner, n))]$

Proof. Let σ and $n \mathring{>} \mathring{0}$ be given and suppose $R(\ulcorner\sigma\urcorner, n)$. We prove this theorem via induction on the complexity of σ. If σ is an atomic truth-free formula, then $R(\ulcorner\sigma\urcorner, \mathring{1})$ and we know that $Tr(\ulcorner\sigma\urcorner, 1) \leftrightarrow Tr_{At}(\ulcorner\sigma\urcorner)$ by ATT1. We know that Tr_{At} is materially adequate for atomic truth-free formulas, and thus $Tr(\ulcorner\sigma\urcorner, 1) \leftrightarrow \sigma$. We now suppose for induction that the theorem holds for all subformulae of σ.

- If σ is of the form $\alpha \wedge \beta$ then we know that $Tr(\ulcorner\alpha \wedge \beta\urcorner, n) \leftrightarrow Tr(\ulcorner\alpha\urcorner, n) \wedge Tr(\ulcorner\beta\urcorner, n)$ by ATT2. We know by inductive hypothesis that that $\alpha \leftrightarrow Tr(\ulcorner\alpha\urcorner, n)$ and $\beta \leftrightarrow Tr(\ulcorner\beta\urcorner, n)$ and therefore $\sigma \leftrightarrow Tr(\ulcorner\sigma\urcorner, n)$.

- The case for σ of the form $\alpha \vee \beta$ is similar and omitted.

- Suppose σ is of the form $\neg\tau$. We then have, by the axiom of ATT4, that $Tr(\ulcorner\neg\tau\urcorner, n) \leftrightarrow \neg Tr(\ulcorner\tau\urcorner, n)$. By inductive hypothesis and logic we know that $\neg\tau \leftrightarrow \neg Tr(\ulcorner\tau\urcorner, n)$ and hence we are done.

- Suppose σ is of the form $\forall x\varphi(x)$. By the axiom ATT5 we know that $Tr(\ulcorner\forall x\varphi(x)\urcorner, n) \leftrightarrow \forall a\, Tr(\ulcorner\varphi([x/a])\urcorner, n)$. We also know,, by inductive hypothesis, that $\varphi(a) \leftrightarrow Tr(\ulcorner\varphi([x/a])\urcorner, n)$ for every a and thus $\forall x\varphi(x) \leftrightarrow \forall a\, Tr(\ulcorner\varphi[x/a])$ and hence we are done.
- The case for σ of the form $\exists x\varphi(x)$ is similar and omitted.
- Finally if σ is of the form $Tr(\ulcorner\tau\urcorner, n-1)$ then we the statement holds by ATT7 directly.

□

We can actually improve this result for the T-Out direction and we get the general rule that $Tr(\ulcorner\sigma\urcorner, n) \rightarrow \sigma$ no matter the rank of σ. This is a quirk of the fact that by ATT10 if σ has rank 0 it will never be the case that $Tr(\ulcorner\sigma\urcorner, n)$, and thus the conditional holds trivially.

Corollary 5.2.4.3 (T-Out)**.** *For any $\mathscr{L}_{\mathrm{Tr}}$-formula σ:*

$$\mathrm{B} + \mathrm{ATT} \vdash \forall n[\mathrm{Tr}(\ulcorner\sigma\urcorner, n) \rightarrow \sigma]$$

The theorem has another important corollary which is that truth respects logical equivalence for sentences of non-zero rank. If two sentences are logically equivalent, then their alethic status is equivalent as well.

Corollary 5.2.4.4 (Classicality)**.** *For any $\mathscr{L}_{\mathrm{Tr}}$-formulas φ and ψ:*

$$\mathrm{B} + \mathrm{ATT} \vdash \forall n[(R(\ulcorner\varphi\urcorner, n) \wedge R(\ulcorner\psi\urcorner, n) \wedge n \mathring{>} \mathring{0}) \rightarrow \\ \rightarrow ((\varphi \leftrightarrow \psi) \leftrightarrow (\mathrm{Tr}(\ulcorner\varphi\urcorner, n) \leftrightarrow \mathrm{Tr}(\ulcorner\psi\urcorner, n)))]$$

Proof. Let φ and ψ be given and suppose $R(\ulcorner\varphi\urcorner, n)$, $R(\ulcorner\psi\urcorner, n)$ and $n \mathring{>} \mathring{0}$. By Theorem 5.2.4.2 we have that $(\varphi \leftrightarrow \psi) \leftrightarrow Tr(\ulcorner\varphi \leftrightarrow \psi\urcorner, n)$. Applying Lemma 5.2.3.2 we hence have that $(\varphi \leftrightarrow \psi) \leftrightarrow (Tr(\ulcorner\varphi\urcorner, n) \leftrightarrow Tr(\ulcorner\psi\urcorner, n))$. □

A highly useful benefit of this corollary is that classical logic is emulated within the truth predicate for all sentences of non-zero rank. This means that classical equivalences such as De Morgan's laws hold internally, as well as rules like double negation elimination. This allows external classical reasoning, as used by our base theory B, to be emulated and respected by the truth predicate. This also enables proof of the following theorem, which was teased in Section 5.2.2. I stated that a formula with rank $a > 1$ is equivalent to a formula, also of rank a, of the form $Tr(x, a-1)$. This is now stated and proven precisely, using the above Corollary 5.2.4.4 and the axioms of ATT.

Lemma 5.2.4.5. *Let σ be an $\mathscr{L}_{\mathrm{Tr}}$-formula and suppose $R(\sigma, a)$ where $a \mathring{>} \mathring{1}$. There is a formula φ such that* $\mathrm{B}+\mathrm{ATT} \vdash \sigma \leftrightarrow \varphi$ *and φ is a formula of the form* $\mathrm{Tr}(x, b)$*, where $R(x, b)$ and $b \mathring{=} (a \mathring{-} 1)$.*[2] *Further,* $\mathrm{B} + \mathrm{ATT} \vdash \mathrm{Tr}(\sigma, a) \leftrightarrow \mathrm{Tr}(\varphi, a)$.

Proof. We prove this lemma via induction on the complexity of σ. If σ is an atomic formula, then σ is of the form $Tr(x, b)$ for some $b \mathring{<} a$. We hence take φ to be $Tr(x, (a \mathring{-} 1))$ and we are done.

- Suppose σ is of the form $\neg\eta$. We know by induction that there is a formula $Tr(y, (a \mathring{-} 1))$ for which the lemma holds for η and hence take φ to be $Tr(\dot{\neg} y, (a \mathring{-} 1))$. We know by R2 $R(\dot{\neg} y, (a \mathring{-} 1))$ and by ATT4 that $\sigma \leftrightarrow \varphi$.

- Suppose σ is of the form $\alpha \wedge \beta$. We then know by induction that there are formulas $Tr(y, (a \mathring{-} 1))$ and $Tr(z, (a \mathring{-} 1))$ for which the lemma holds for α and β respectively. We take φ to be $Tr(y \dot{\wedge} z, (a \mathring{-} 1))$. We know by R3 that $R(y \dot{\wedge} z, (a \mathring{-} 1))$ and by ATT2 that $\sigma \leftrightarrow \varphi$.

- The case for σ of the form $\alpha \vee \beta$ is similar and omitted.

- Suppose σ is of the form $\forall x \eta(x)$. We then know by induction that there is a formula $Tr(y(i), (a \mathring{-} 1))$ for which the lemma holds for $\eta(i)$ for each i. We take φ to be $Tr(\forall x y(x), (a \mathring{-} 1))$. We know by R5 that $R(\forall x y(x), (a \mathring{-} 1))$ and by ATT5 that $\sigma \leftrightarrow \varphi$.

- The case for σ of the form $\exists x \eta(x)$ is similar and omitted.

This shows the first claim. For the second claim we note that this follows directly from Corollary 5.2.4.4 since $a \mathring{>} \mathring{1}$. □

This lemma is useful, as it allows us to prove the connection between $\mathrm{PA}+\mathrm{ATT}$ and $\mathrm{PA} + \mathrm{RT}_{\epsilon_0}$. Both theories believe that the same sentences are true, given an appropriate translation between the two languages. We interpret σ of the form $Tr^n(\varphi)$ in $\mathscr{L}_{\mathrm{RT}_{<\epsilon_0}}$ as $Tr(\varphi, n)$ in $\mathscr{L}_{Tr} \backslash \{R\}$ and vice-versa.

Theorem 5.2.4.6. *For each $n < \epsilon_0$ and φ in $\mathscr{L}_{\mathrm{RT}_{<\epsilon_0}}$:*

$$\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash \mathrm{Tr}^n(\varphi) \text{ if and only if } \mathrm{PA} + \mathrm{ATT} \vdash \mathrm{Tr}(\varphi, n)$$

[2]Here we are treating $(a \mathring{-} 1)$ as a shorthand for the code of an ordinal which is equal to the result of subtracting 1 from the ordinal that a codes.

Proof. We prove this via induction on n. First we assume $n = 1$. If $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash Tr^1(\varphi)$ then φ is an $\mathscr{L}_A$ formula and hence $\mathrm{PA} + \mathrm{ATT} \vdash R(\varphi, 1)$ by Lemma 5.2.2.1. We note that the axioms of ATT restricted only to formulae of rank 1 are identical to the axioms of $\mathrm{PA} + \mathrm{RT}_1$ and hence $\mathrm{PA} + \mathrm{ATT} \vdash Tr(\varphi, 1)$. For the converse we note that if $\mathrm{PA} + \mathrm{ATT} \vdash Tr(\varphi, 1)$ then by ATT9 $R(\varphi, 1)$ and thus again by Lemma 5.2.2.1 we know φ is an $\mathscr{L}_A$ formula. Therefore $\mathrm{PA} + \mathrm{RT}_1 \vdash Tr^1(\varphi)$.

We now assume the theorem holds for all $k < n$. If $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash Tr^n(\varphi)$ then either $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash Tr^{n-1}(\varphi)$ or using a result due to Cieśliński (2010b, p. 334) φ is equivalent to a formula of the form $Tr^{n-1}(\delta)$. In the former case we are done by induction, so thus we assume φ is equivalent to a formula of the form $Tr^{n-1}(\delta)$. We know $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0}$ is materially adequate and hence $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash \varphi$, i.e. $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash Tr^{n-1}(\delta)$. We now use our inductive hypothesis to deduce that $\mathrm{PA} + \mathrm{ATT} \vdash Tr(\delta, n-1)$ and hence by ATT7 $\mathrm{PA} + \mathrm{ATT} \vdash Tr(Tr(\delta, n-1), n)$. We now use Lemma 5.2.4.5 to deduce that $\mathrm{PA} + \mathrm{ATT} \vdash Tr(\varphi, n)$. For the converse direction we can follow these steps in reverse. If $\mathrm{PA} + \mathrm{ATT} \vdash Tr(\varphi, n)$ then either $\mathrm{PA} + \mathrm{ATT} \vdash Tr(\varphi, n-1)$ and we are done by induction or $R(\varphi, n)$ and by Lemma 5.2.4.5 φ is equivalent to a formula δ of the form $Tr(\sigma, n-1)$. By Corollary 5.2.4.3 we know $\mathrm{PA} + \mathrm{ATT} \vdash Tr(\sigma, n-1)$ and by inductive hypothesis $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash Tr^{n-1}(\sigma)$. We now use the material adequacy of $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0}$ to deduce $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash Tr^n(\delta)$ and hence by classicality $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash Tr^n(\varphi)$. □

This result gives us a lower bound on the arithmetic strength of PA+ATT. This follows from a result due to Feferman (1991, §4) where he shows that $\mathrm{PA} + \mathrm{RT}_{\leqslant n}$ and $\mathrm{RA}_{\leqslant n}$[3] are intertranslatable.

Corollary 5.2.4.7. $\mathrm{PA} + \mathrm{ATT}$ *can prove all the arithmetical consequences of* $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0}$ *and hence the arithmetic strength of* $\mathrm{PA} + \mathrm{ATT}$ *is bounded below by* $\mathrm{RA}_{<\epsilon_0}$.

Proof. We use the translation given above. If $\mathrm{PA} + \mathrm{RT}_{<\epsilon_0} \vdash \varphi$, then by material adequacy $\mathrm{PA}+\mathrm{RT}_{<\epsilon_0} \vdash Tr^n(\varphi)$ for an appropriate n. Therefore by Theorem 5.2.4.6 $\mathrm{PA} + \mathrm{ATT} \vdash Tr(\varphi, n)$ and hence by Corollary 5.2.4.3 $\mathrm{PA} + \mathrm{ATT} \vdash \varphi$ as well. For the second part of the Corollary, we note that this follows directly from a result due to Feferman (1991, p. 18) that $\mathrm{PA} + \mathrm{RT}_{<\alpha}$ and $\mathrm{RA}_{<\alpha}$ are interdefinable. □

[3]For details on the systems of ramified analysis $\mathrm{RA}_{\leqslant n}$, the reader is referred to Feferman's *Systems of Predicative Analysis* (Feferman, 1964, p. 21-22).

It appears reasonable to weakly conjecture that the arithmetic strength of PA + ATT is exactly $\mathrm{RA}_{<\epsilon_0}$, since the additional provability given by ATT is solely about the truth predicate, although this remains an open question.

Conjecture 5.2.4.8. *The arithmetic strength of* PA + ATT *is exactly* $\mathrm{RA}_{<\epsilon_0}$

This result also has the following important corollary, relevant to Chapter 2. We know that PA + ATT can prove the alethic consequences of CT (for an appropriate translation between the languages) and therefore PA + ATT can also prove all instances of the (typed) extended T-Schema from Chapter 2, Definition 2.5.2.

Corollary 5.2.4.9. $\mathrm{PA}+\mathrm{CT}^-+\vDash^*$ *is a subtheory of* PA+ATT*, for an appropriate translation between the languages.*

It is also interesting to consider what the strength of PA + ATT^- is, where ATT^- is ATT without any induction axioms. In particular, it is interesting to see whether this theory is conservative (as PA + CT^- is) or not given interest in the 'conservativity argument' against deflationism discussed in Chapters 2 and 4.

Question 5.2.4.10. *Is* PA + ATT^- *proof-theoretically conservative over* PA*?*

Leaving open questions aside for now, there are two basic intuitions on truth that it is desirable for any truth theory to exhibit: that the predicate is internally consistent (no sentence is both true and its negation true) and complete (every 'truth-apt' sentence is true or its negation is true). The first intuition holds directly in ATT.

Lemma 5.2.4.11 (Consistency)**.** $\mathrm{B} + \mathrm{ATT} \vdash \forall\sigma\forall n[\neg(\mathrm{Tr}(\sigma, n) \wedge \mathrm{Tr}(\neg\sigma, n))]$.

Proof. Suppose for contradiction there are σ and n such that $Tr(\sigma, n) \wedge Tr(\neg\sigma, n)$. By Axiom ATT2 we deduce $Tr(\sigma \dot{\wedge} \neg\sigma, n)$. Thus by Corollary 5.2.4.3 we conclude $\sigma \wedge \neg\sigma$ which is a contradiction. □

The intuition that truth is complete is not as precisely formulated as the the intuition that truth is consistent. There are many sentences of natural language for which it appears that neither they nor their negation are true: sentences which (informally) do not express any truth-apt semantic content. These include imperatives such as 'Do not run in the corridor' or exclamations such as 'Boo!'. These

sentences are not true, but it would seem strange to say their negations 'Do run in the corridor' and 'not boo!' are also true. Such sentences are called not truth-apt, and this is my preferred interpretation of formulas which have rank 0. This will be discussed in more detail in Section 5.2.5, but for now it suffices to remark that it appears that every truth-apt formula or its negation should be true. This is precisely the form of completeness that ATT exhibits given this interpretation: every sentence with non-zero rank is true or its negation is true.

Lemma 5.2.4.12 (Completeness)**.** $\mathrm{B} + \mathrm{ATT} \vdash \forall\sigma\forall n[(R(\sigma, n) \wedge n \mathring{>} \mathring{0}) \rightarrow$
$\rightarrow (\mathrm{Tr}(\sigma, n) \vee \mathrm{Tr}(\neg\sigma, n))]$

Proof. Let σ and n be given and suppose $R(\sigma, n)$ where $n \mathring{>} \mathring{0}$. If $Tr(\sigma, n)$ then we are done, hence we suppose $\neg\, Tr(\sigma, n)$. By ATT4 we have $Tr(\neg\sigma, n)$. □

This section hence shows that a number of important and desirable properties for truth hold for all sentences of non-zero rank. We get compositionality, an extended T-Schema from Chapter 2, and internal consistency and completeness. This should demonstrate that, certainly with respect to sentences of non-zero rank, ATT is formally adequate in the sense of Chapter 4. ATT can provide for all the syntactic features we desire of a truth predicate without losing consistency. This, of course, comes with the important caveat that these features do not tend to hold for sentences with rank 0. In the following section these sentences will be discussed and explored in more detail.

5.2.5 ATT and Rank 0

Given the theorems above, ATT satisfies many desirable features of a theory of truth for sentences of non-zero rank. The theory behaves classically, is internally consistent and complete, is compositional and satisfies both the T-In and T-Out rules. These properties are all part of our basic intuitions of how a theory of truth should behave and I believe show formal adequacy for ATT with respect to non-zero rank sentence in the sense of Chapter 4. The reason that these norms are jointly satisfied by ATT, without inconsistency, is that they are restricted to only sentences of non-zero rank. It is important to note that this is the vast majority of sentences of $\mathscr{L}_{Tr}$ (in particular, all sentences of $\mathscr{L}$ as well as most of those involving the truth predicate). In this section we now consider the sentences

which do not conform and have rank 0 instead. We know by ATT10 that these sentences are not true, and that their negation is also not true.

When we look at how sentences gain rank, we see that 'ordinary' sentences will have rank greater than 0. If a sentence belongs to the base language $\mathscr{L}$, then it will always have rank 1 by Lemma 5.2.2.1. If logical connectives or quantifiers are applied to this sentence, the rank will never decrease, and if the truth predicate is applied to the sentence, then it will increase its rank by 1, remaining above 0. Direct uses of the truth predicate on sentences of the base language, or already settled alethic sentences, provides us with a sentence of non-zero rank, and an attractive typed treatment of a sentence's alethic status. The rank 0 sentences cannot arise this way, and instead fall into two general categories: paradoxes and sentences which quantify absolutely. Paradoxes are classified here as uses of the truth predicate on unsettled (and unsettlable) alethic sentences, and sentences which quantify absolutely are those which quantify over all levels of the truth predicate. Both of these cases will be discussed in this section.

Once a sentence of rank 0 has been produced, a family of such sentences can be generated. Negating or adding a quantifier to a formula with rank 0 results in the new formula also possessing rank 0. By taking its disjunction or conjunction with itself, we similarly get a new rank 0 formula. More interestingly, adding a conjunct or disjunct to a rank 0 formula, where the other clause has rank $n > 0$, results in the new sentence having rank n also, due to R2). If this sentence is a conjunction, then it shall never be true, but if it is a disjunction, then this will be true if the other disjunct is true. This follows directly from the axioms for ATT, and means that a sentence such as $\sigma \vee 0 = 0$ is always true at any level $n > 0$ even if σ has rank 0. This is quite an attractive feature of the theory, and allows consistent generation of as many positive-rank formulas as possible. We will discuss a three-valued logic featuring this behaviour and its use for developing a type-free theory of truth in much more detail in Section 5.3.

My chosen philosophical interpretation of the rank 0 sentences is that these should be deemed as not truth-apt. We are working within the metatheory of classical logic, and only have two truth-values: true (at some level) and false (at all levels). The rank 0 formulas are not true at any level, but nor are they false at all levels either, since their negations are also not true at any level. These sentences do not fall in the extension of anti-extension of our truth-predicate, but are still provable or refutable in our theory and carry expressible meaningful content. This

does not mean that these sentences are not well-formed, or unknown, but they are not the sort of things that we should predicate truth of. These sentences are not truth-apt, in the sense that the truth predicate cannot meaningfully apply to these sentences.

There are many natural language sentences which fall into this category: they are well formed, expressible and meaningful, but we should not apply the truth predicate to these sentences. An imperative, such as "read this thesis!" or its negation "do not read this thesis!", is one type of sentence which is not truth-apt. It does not make sense to say that either of these commands are true, despite one being a negation of the other. Questions are similarly not truth-apt, such as: "should I read this thesis?" and its negation "should I not read this thesis?" It would be very strange to say that a question is true.

ATT is of course not a theory dealing with imperatives or questions, but does suggest two other examples of non-truth-apt sentences. These are the rank 0 sentences, which fail to be true at any level, and also fail to have a true negation. The rank 0 sentences are those which unrestrictedly quantify over levels of the truth predicate and sentences which are paradoxical in nature, such as the Liar paradox. Both are philosophically problematic notions, and to find that they are given a diagnosis and treatment by ATT is an attractive facet of the theory. In the following section I consider the sentences which quantify over absolutely all levels of the truth predicate in more detail, and the connection that deeming these as not-truth-apt has to debates about absolute generality in philosophy of logic.

Unrestricted Quantification of Truth Predicates

One method of producing rank 0 formulas is by quantifying over absolutely all levels of the truth-predicate. For example, a formula such as:

$$(\star) \qquad \forall x[(Ord(x) \wedge x \mathring{>} \mathring{0}) \rightarrow Tr(\sigma, x)]$$

has rank 0, as well as a formula such as $\exists y[Ord(y) \wedge Tr(\delta, y)]$. These formulas quantify over absolutely all levels of the truth predicate and tell us something about the behaviour of σ and δ across the entire typed hierarchy of truth.

The formal reason these sentences are given rank 0 follows from R5 that $R(\star, k)$ if and only if $k = sup\{y_i : R((Ord(y_i) \wedge y_i \mathring{>} \mathring{0}) \rightarrow Tr(\sigma, i), y_i)\}$. We know by R7 that hence $k \geqslant sup\{y_i : Ord(y_i)\}$ and thus as there is no maximal ordinal, by our

agreement on the formalisation of the supremum function, $k = 0$.

It should be noted that although these sentences which quantify over all levels of truth has rank 0, any substitutional instance of these sentences is perfectly admissible. For instance the sentence $(Ord(n) \wedge n \mathring{>} \mathring{0}) \rightarrow Tr(\sigma, n)$ has rank n and can be proven true or false at the level $n + 1$. As Russell remarks, for a typed theory there is an important distinction between 'any' and 'all'.

> In the case of such variables as propositions or properties, 'any value' is legitimate, though 'all values' is not (Russell, 1908, p. 229).

Whilst any of the instances of $\star$ is truth-apt, the sentence $\star$ which expresses all instances is not.

This tells us that whilst we can prove claims like $\star$ and $\neg\exists x\, Tr(\ulcorner 0 = 1 \urcorner, x)$ internally within ATT, without recourse to the metatheory, we cannot prove that these claims will ever become true at a certain level. Were one of these sentences to become true at a certain level, then it would be making alethic claims about levels beyond itself, and contradict the typed nature of truth in ATT (and in particular axioms R6 and T9). Whilst our theory can quantify over absolutely all levels of the truth predicate (for our particular language), in doing so we produce a sentence which itself fails to be either true or false. These absolutely general sentences are well-formed, and even provable within the theory, but under my interpretation are not truth-apt. It is a strength of the theory that it is able to internally prove sentences about all levels of the truth predicate, and this sets it apart from a typed theory of truth like $\mathrm{RT}_{<\lambda}$ for an ordinal λ. It is a necessary feature of the theory that such sentences cannot be proven true at a particular level, and this follows from its typed nature. This shows that proof and truth come apart for ATT and we cannot have the general scheme: $\mathrm{B} + \mathrm{ATT} \vdash \sigma \rightarrow \exists x\, Tr(\ulcorner \sigma \urcorner, x)$ without inconsistency.

Parsons (1974) argues for a position which is very similar to ATT's behaviour in this regard. Parsons considers the Liar paradox and the paradoxes of set theory and argues that these imply certain well-formed formulae nevertheless fail to be truth-apt. The former shows that the Liar sentence fails to express a proposition, and the latter show that certain (meaningful) predicates fail to express a set. Parsons concludes:

> A language may contain perfectly meaningful predicates such that, in a given theory formulated in that language, they cannot be said to

> have extensions ... the same situation arises for *sentences*: A theory expressed in a given language cannot always correlate to a sentence a proposition as its intension, even though the sentence is well-formed and may even be provable (Parsons, 1974, p. 390).

This behaviour means that ATT has a stake in current debates over whether it is possible for quantifiers to quantify over 'absolutely everything'. If one wishes to take ATT as a theory which approximately describes our natural-language concept of truth, then natural language quantifiers never range over absolutely everything, since they cannot range over absolutely all of the truth hierarchy.[4] ATT proves that it is not possible for a sentence to quantify over all levels of the truth predicate and remain truth-apt, and thus not everything can be quantified over. That ATT implies that not absolutely everything can be quantified over, does not lead to an unorthodox stance.

It is certainly unclear whether it is legitimate to quantify over absolutely everything, or if there could be, as Rayo and Uzquiano (2006, p. 2) ask, "an all-inclusive domain be available to us as a domain of inquiry?". Hellman (2006), for example, details and supports one of the main arguments against our ability to quantify over absolutely everything. This is the argument from indefinite extensibility: some mathematical concepts appear to have no limit – such as the concept of an ordinal number. Given any precise collection of ordinals, we are able to consider the limit of these, and conceive of a new ordinal outside of our collection. We hence cannot quantify over absolutely all ordinals (and thus absolutely everything), since to do so we would be forced to conceive of a new ordinal which we had not quantified over. This situation is very similar to ATT's behaviour, particularly if we imagine ATT as a general theory of truth for natural language. A sentence which quantifies over all levels of the truth predicate quantifies over all ordinals, but as Hellman argues this is not possible. For a generalised form of ATT, the hierarchy of truth is an indefinitely extensible concept.

Given ATT's similarity with a contextual approach to truth, it is pleasing that this behaviour is in harmony with the contextualist position within the debate on absolute generality too. Glanzberg (2004) champions a contextual approach to truth and quantification, and provides an alternative argument against our ability to quantify over absolutely everything. Glanzberg argues that all uses of

[4]It should be noted that even expressing this position in its intended spirit, without apparent contradiction, is no easy matter. Williamson (2003, §5) provides a lively discussion of this.

quantifiers in natural language can be indefinitely extensible, and thus it is not possible to determinately[5] specify a maximal domain of 'absolutely everything' due to a general form of Russell's paradox. Glanzberg champions a contextual interpretation of this and advances the view that quantifiers are context-dependent: a shift in context can always widen (or shrink) the domain that they quantify over. In ATT the domain of the hierarchies of the truth predicate can always be widened, and there is similarly no maximal domain to be quantified over. If the levels of the truth predicate are interpreted as contexts, then it is concordant (albeit not necessary) to interpret the quantifiers as similarly contextual.

I do not mean to imply that the debate over absolute generality is settled, or that these arguments have not been critiqued. Williamson (2003) defends the position that we can quantify over absolutely everything, for example. Yet it is certainly a defensible feature of ATT, with interesting philosophical consequences, that a sentence in $\mathscr{L}_{Tr}$ containing quantification over all levels has rank 0 and is not truth-apt. It shows that ATT has philosophical consequences and behaviour which extends beyond questions over truth, and suggests further research into connections between truth and absolute quantification. In the next section I move away from absolute quantification and consider the other type of rank 0 formulas: paradoxes.

Paradoxes

The other type of formulas that ATT classifies as rank 0 are those which are paradoxical and utilise the truth predicate to produce a formula which is 'ill-founded' in some sense. These examples are most commonly seen in the traditional semantic paradoxes. We will not explicitly construct such formulas here, but it is an easy exercise to construct them using a diagonalisation lemma. For the details on this see a standard textbook on the subject, such as Kaye's (1991, Lemma 3.8) *Models of Peano Arithmetic*. In this section we will explore various different formulations of the Liar paradox, and how ATT responds to these. The first sentence that we shall examine is a Liar-like sentence that says of itself it is never true at any level.

$$\text{Unbounded Liar } \lambda: \qquad \lambda \leftrightarrow \forall n[\neg Tr(\ulcorner \lambda \urcorner, n)]$$

[5]Glanzberg defines this here to mean a domain which sharply and exhaustively tells us everything that is in the domain

This sentence says of itself that it is not true at any level of the truth predicate. Since λ quantifies over absolutely all levels of the truth predicate it also fits into the category above and thus we immediately see that it has rank 0. Since the rank 0 sentences are never true, we know that $\mathrm{B} + \mathrm{ATT} \vdash \forall n[\neg Tr(\ulcorner \lambda \urcorner, n)]$, and hence $\mathrm{B} + \mathrm{ATT} \vdash \lambda$. This approach to paradox tell us that the Liar sentence is provable, and provably not true.

That ATT proves the Liar and proves that the Liar is not true is not formally inconsistent, since the theory admits that there are sentences which are not truth-apt, but can still be provable. As stated above, ATT denies that proof necessarily implies truth. The theory (formally) avoids any revenge-style paradox by blocking any alethic reasoning about a rank 0 sentence, other than to say that it is never true. Just because λ states that this is the case, is not enough to make the sentence internally true, since it is deemed to be a problematic sentence for which ordinary truth-reasoning (such as T-In) does not apply.

This response can be seen as ATT's general approach to paradoxes. There are formulas for which our ordinary alethic reasoning does not apply (the rank 0 formulas) and these should be considered as untrue, despite what they say. Sometimes this means that these formulas are provable, and sometimes they are refutable, but being provable is not a sufficient condition for truth. This is similar to Glanzberg's (2001) contextual approach to truth, and whilst this is not uncontroversial by any means, it seems that his arguments can be adapted for ATT's defence as well. Whether such a response is philosophically desirable is widely debated, and one I shall set aside here, to focus on ATT's formal approach to paradox.

A different Liar-style sentence is one which says that its negation is true (at a level). I will call this style of Liar sentence an F-Liar (for falsity) and first consider its quantified case.

$$\text{Unbounded F-Liar } \lambda^F\text{:} \qquad \lambda^F \leftrightarrow \exists n[\,Tr(\ulcorner \neg \lambda^F \urcorner, n)]$$

This sentence also has rank 0, but unlike the Liar above, it is refutable. We know that there is no level n where $\neg\lambda^F$ is true (since $\neg\lambda^F$ also has rank 0 by R2, and thus will never be true by ATT10). Since this is the negation of what λ^F says, we know that therefore $\neg\lambda^F$ is provable.

These examples of paradoxes show how proof and truth pull apart for rank 0 sentences, but it might be thought not truly how ATT deals with paradox. I have

shown that these sentences have rank 0 because they quantify over absolutely all levels of the truth predicate, as per the examples in Subsection 5.2.5. We can instead consider paradoxical sentences which refer only to a particular level, and also naively result in contradiction. These are not paradoxical because of their scope of quantification, but because of their self-referential nature. It is illuminating to see how the typed nature of ATT deals with these sentences. First I will consider a family of F-Liars, which say of themselves that their negation is true at a particular level.

$$\text{Level } n \text{ F-Liar } \lambda_n^F\text{:} \qquad \lambda_n^F \leftrightarrow Tr(\ulcorner \neg \lambda_n^F \urcorner, \mathring{n})$$

We can prove that these sentences are also of rank 0, even though they do not contain any quantifiers. This is a prime example of how paradoxes can also generate 'not truth-apt' sentences.

Lemma 5.2.5.1. *For each n:* $\mathrm{B} + \mathrm{ATT} \vdash R(\lambda_n^F, \mathring{0}) \wedge \neg\lambda_n^F$

Proof. Fix $\mathscr{M} \vDash \mathrm{B} + \mathrm{ATT}$ and $k \in dom(\mathscr{M})$ such that $\mathscr{M} \vDash R(\lambda_n^F, k)$ and suppose for contradiction that $k \mathring{>} \mathring{0}$. We work inside $\mathscr{M}$ and know by Lemma 5.2.4.12 that $Tr(\ulcorner \lambda_n^F \urcorner, k) \vee Tr(\ulcorner \neg \lambda_n^F \urcorner, k)$.

If $Tr(\ulcorner \lambda_n^F \urcorner, k)$ then we know that $\neg Tr(\ulcorner \neg \lambda_n^F \urcorner, n)$. This is because if $k > n$, then $\neg Tr(\ulcorner \neg \lambda_n^F \urcorner, n)$ by ATT9, or if $k \leqslant n$ then, using Lemma 5.2.4.11 and ATT8, also $\neg Tr(\ulcorner \neg \lambda_n^F \urcorner, n)$. We thus deduce $\neg \lambda_n^F$ by logic, but by Theorem 5.2.4.2 we also have λ_n^F and thus a contradiction.

Therefore we derive $Tr(\ulcorner \neg \lambda_n^F \urcorner, k)$ and hence λ_n^F by ATT8 and the definition of λ_n^F. We also derive $\neg \lambda_n^F$ by Theorem 5.2.4.2 and hence we also have a contradiction. Therefore $k \mathring{=} \mathring{0}$ and also $R(\ulcorner \neg \lambda_n^F \urcorner, \mathring{0})$, so $\neg Tr(\ulcorner \neg \lambda_n^F \urcorner, n)$ by ATT10. Thus we also derive $\neg \lambda_n^F$. □

We can also consider the family of Liar-like sentences which say of themselves that they are not true at a particular level. These sentences, analogous to the unbounded Liar, are also provable and provably not true at the level they specify.

$$\text{Level } n \text{ Liar } \lambda_n\text{:} \qquad \lambda_n \leftrightarrow \neg Tr(\ulcorner \lambda_n \urcorner, \mathring{n})$$

Unlike the F-Liar sentences above, the level n Liar Paradoxes cannot be proven to have rank 0, although the canonical interpretation of the rank predicate will say this. There are models of ATT where these sentences instead have a rank

strictly larger than n. This is because these sentences are also never true at rank n by ATT9 and hence also provable and provably untrue at rank n without contradiction.

Lemma 5.2.5.2. *For each n:* $\mathrm{B} + \mathrm{ATT} \vdash R(\lambda_n, k) \wedge (k \mathring{>} \mathring{n} \vee k \mathring{=} \mathring{0}) \wedge \lambda_n$

Proof. Fix $\mathscr{M} \vDash \mathrm{B} + \mathrm{ATT}$ and $k \in dom(\mathscr{M})$ such that $\mathscr{M} \vDash R(\lambda_n, k)$ and suppose for contradiction that $\mathring{n} \mathring{\geqslant} k \mathring{>} \mathring{0}$. We work inside $\mathscr{M}$ and suppose λ_n. We deduce $Tr(\ulcorner \lambda_n \urcorner, k)$ by Theorem 5.2.4.2 and hence $Tr(\ulcorner \lambda_n \urcorner, n)$ by ATT8, which implies $\neg\lambda$ and thus a contradiction. We hence suppose $\neg\lambda_n$ and equivalently $Tr(\ulcorner \lambda_n \urcorner, \mathring{n})$. We again use Theorem 5.2.4.2 to derive λ_n, which is also a contradiction. Therefore $k \mathring{=} \mathring{0} \vee k \mathring{>} \mathring{n}$ and hence $\neg\, Tr(\ulcorner \lambda_n \urcorner, \mathring{n}) \wedge \lambda_n$. □

A sentence often thought of as similar, if far less problematic, than the Liar Paradox is the truth-teller sentence τ which says of itself that it is true. Unlike the Liar paradox, this doesn't seem contradictory, but does seem to make evaluation upon whether it is actually true or false impossible. As with all sentences which quantify over absolutely all levels of the truth predicate, the version of the truth-teller which says there is some level at which it is true has rank 0 and thus will never be true. More formally, we define this unbounded truth-teller in the following way.

$$\text{Unbounded Truth-Teller } \tau\text{:} \qquad \tau \leftrightarrow \exists n\, Tr(\ulcorner \tau \urcorner, n)$$

Since this sentences says that there is a level at which it is true, but there is no such level at which it is true (since it has rank 0 and is always false) $\neg\tau$ is provable. This means that ATT can provide information on some forms of the truth-teller, an attractive feature for those interested in what the theory can say about other semantic puzzles. We can also consider a family of more traditional truth-tellers which say of themselves that they are true at a particular level n.

$$\text{Level } n \text{ Truth-teller } \tau_n\text{:} \qquad \tau_n \leftrightarrow Tr(\ulcorner \tau_n \urcorner, \mathring{n})$$

These sentences are neither provable nor refutable by ATT, since there are models of both $\mathrm{B} + \mathrm{ATT} + \tau_n$ and $\mathrm{B} + \mathrm{ATT} + \neg\tau_n$ for each $n > 0$. This meets intuitions that the truth-teller does not carry enough semantic information within itself to be determined true or false, although as ATT is a classical theory it will still prove $\tau_n \vee \neg\tau_n$ for each n. The level n truth-tellers are so undetermined that

we cannot even say what rank they have, although the canonical rank predicate will give them rank 0.

These remarks show how ATT approaches liar paradoxes in various guises as well as truth-teller sentences too. One other form of semantic paradox of interest is the Yablo-Visser paradox (Yablo, 1985; Visser, 1989), which is an infinitely long list of sentences each of which states that every sentence appearing later in the sequence is untrue. It is an interesting open question to consider how ATT approaches and deals with these sentences.

Question 5.2.5.3. *Can we formulate Yablo-Visser style paradoxes in* $\mathscr{L}_{\mathrm{Tr}}$*, and if so, how does* ATT *deal with these?*

I leave this question open as an area for further research. I hope that this subsection has sufficiently shown there is great interest in the behaviour of ATT's rank 0 sentences, and I hope that this section has shown the interesting features of ATT as a theory of truth more generally. I particularly hope that this section has shown that ATT's behaviour with respect to the rank 0 sentences is defensible. I hope that this shows that whilst ATT does not universally satisfy all we might hope for from a theory of truth (as no theory can), where it falls short has a philosophical explanation. I hope this defends ATT's formal adequacy in the sense of Chapter 4. In the following section I move away from ATT to discuss a new type-free theory of truth that results from reflecting upon the rank of sentences in ATT.

5.3 KFJ

In the previous section I defined and discussed a theory of axiomatic typed truth (ATT); a theory which classifies sentences of its languages by their alethic rank and speaks of a sentence's truth value only relative to a particular level within ATT's truth-hierarchy. I showed that the theory's behaviour for positive-rank formulas is pre-theoretically highly desirable, and then discussed in some detail the sentences of rank 0. These sentences of rank 0 are classified as untrue at any level, as are their negations.

The theory of ATT is inherently typed, but can be used to develop a type-free theory of truth. We could introduce a type-free truth property T for the language $\mathscr{L}_{Tr}$ which satisfies the schema $\mathrm{B} + \mathrm{ATT} \vdash \exists x\, Tr(\varphi, x) \rightarrow T(\varphi)$. This generates a type-free notion of truth from our typed theory of truth. In this section I will

develop a type-free theory of truth that satisfies this constraint. Rather than building this theory on top of ATT, however, I shall construct it independently, and then prove that this schema is satisfied for an appropriate translation between the languages.

In this section I will construct this KF-like axiomatic theory of truth. The theory is specified axiomatically, but motivated by a novel three-valued logic which can be seen as the internal behaviour of an ATT-interpretation of the $\mathscr{L}_{Tr}$-formula $\exists x\, Tr(\varphi, x)$. This logic is the internal logic of the theory's truth predicate and stands in relation to the theory as Strong Kleene logic stands to KF. It will be shown that this theory can be seen as the 'type-free' variant of ATT and I will remark on the lessons that can be drawn from this. To begin, I introduce and discuss the three-valued logic ML_3 that is the motivating logic of KFJ.

5.3.1 The Internal Logic

To develop this next theory of truth (KFJ), I first introduce a new[6] propositional three valued logic: ML_3. This logic aims to be maximal in the sense that it hopes to maximise the number of classical valuations (those of 0 or 1). As with all three-valued logics, it has three possible semantic valuations $\{-1, 0, 1\}$, which I will also denote as $\{M, 0, 1\}$, and the aim is to maximise the number of sentences which are assigned either 0 or 1, where 1 is our designated truth value and 0 our designated falsity value. The logic tries to avoid assigning values of M (the third semantic value) as much as possible, and the valuation of two connected subformulae will always be 0 or 1 if at least one of these subformulae has a valuation of 0 or 1. This can be thought of as the dual of the weak-Kleene logic, in which a valuation of M to any subformula infects the whole formula and gives it value M as well. In ML_3 we see the opposite behaviour, and a classical subformula is able to bear the brunt of the valuation and give the main formula a classical valuation as well, even if other subformulae have no classical valuation.

Because of this behaviour, my intuitive interpretation of this third value, M or -1, is not 'undefined' as in strong or weak Kleene logic (Kleene, 1952, §64) but instead 'not truth-valued' or 'not truth-apt'.[7] Strictly speaking M should not

[6] As far as I have been able to find, this logic does not exist elsewhere in the literature. The truth-table for $\vee$ in ML_3 is the truth-table for $\vee$ in Sobociński's three-valued logic (Bolc and Borowik, 1992, §3.12), but the similarities end there.

[7] For a more detailed discussion of what I mean by these terms, I refer the reader to Section

be viewed as an alethic value, but as a semantic value, which corresponds with a lack of an alethic value. This is in comparison to 0 and 1, which are intended to correspond with the standard semantic and alethic values of true and false. The aim of ML_3 is for it to be a logic of the classical truth-apt formulas and the formulas which are not truth-apt, and their interaction.

To introduce the logic, I will first provide truth tables for the standard logical connectives. Given a connective $\oplus$ (top lefthand corner) and two formulae i (with semantic valuation in the left-most column) and j (with semantic valuation in the top-most row) we define the valuation of $i \oplus j$ as the intersection of i and j's semantic valuations. I begin with the truth-table for $\vee$:

$\vee$	M	0	1
M	M	0	1
0	0	0	1
1	1	1	1

We see from the table that the only way a disjunction will have valuation M is if both its disjuncts also have valuation M. The logic tries to maximise classical valuations as much as possible by following the (classically correct) rule that 'a disjunction is true if one of its disjuncts is true' and 'a disjunction is false if it is not true and one of the disjuncts is false'. The table for $\wedge$ follows:

$\wedge$	M	0	1
M	M	0	0
0	0	0	0
1	0	0	1

Again we see that the only way a conjunction will have valuation M is if both its conjuncts have valuation M, similarly to disjunction. Here the logic follows the, classically valid, rule that 'a conjunction is true if and only if both its conjuncts are true' and 'a conjunction is false if and only if one of its conjuncts is false'.

The next truth table is for negation and behaves more familiarly:

5.2.5. The distinction appears to be a small one, but I do not understand 'not truth-apt' sentences to be undefined, for they can contain useful content and be provable, but they do not fall into the extension or anti-extension of the truth predicate.

$\neg$	
0	1
M	M
1	0

In ML_3 the connective arrow is the material conditional from classical logic and a formula of the form $A \to B$ is explicitly used as a shorthand for $\neg A \vee B$. I include the table for this connective for completeness, although it can be easily derived from the tables for $\neg$ and $\vee$.

$\to$	M	0	1
M	M	0	1
0	1	1	1
1	0	0	1

The connective for the biconditional is similarly used as a shorthand, and we write a formula of the form $A \leftrightarrow B$ to denote $(A \to B) \wedge (B \to A)$. Again, the table for this connective is included for completeness and could be derived from the tables for $\to$ and $\wedge$. Interestingly, we could equivalently define the biconditional by the rule $A \leftrightarrow B$ is true if and only A and B are true, or A and B are false, and is false if A is true and B is not, A is false and B is not, or vice-versa.

$\leftrightarrow$	M	0	1
M	M	0	0
0	0	1	0
1	0	0	1

These tables define the semantics of the standard propositional connectives in ML_3. We can also give these an algebraic semantics by providing rules of numerical valuation. For these rules I use the valuation space $\{-1, 0, 1\}$ where -1 corresponds to M. I choose -1, rather than $\frac{1}{2}$, as is more common for three-valued logics, for two reasons. Philosophically, it captures the intuition that the third value is 'not truth-valued', rather than 'undefined' (and so below, rather than sitting between, 0 and 1), and pragmatically, it enables the formulation of more elegant rules. As the reader may notice, these are still not always particularly elegant, and relies upon the agreement that $1 - -1 = -1$ in this domain.

- $A \vee B = \max\{A, B\}$

- $A \wedge B = A^2 \times B = A \times B^2$
- $\neg A = 1 - A$
- $A \rightarrow B = \max\{1-A, B\}$
- $A \leftrightarrow B = \max\{1-A, B\}^2 \times \max\{A, 1-B\}$
  $= \max\{1-A, B\} \times \max\{A, 1-B\}^2$

The final remark to make is that ML_3 has the important feature that it is normal, in the sense that it behaves classically for all classical truth-valuations. The less-satisfactory behaviour of ML_3 is that we loose many classical tautologies, such as De Morgan's Laws, and that $P \rightarrow P$.

I have somewhat suggestively written this maximal logic (ML_3) with the subscript 3. This is deliberate, as I leave open the possibility that it can be adapted to a logic with four or more values. This is an open area for further inquiry, as well as the establishment of further results about ML_3.

Question 5.3.1.1. *Are there logics analogous to* ML_3 *which have four or more possible semantic values?*

In the next section I will show how we can use this logic to motivate a KF-style theory of truth.

5.3.2 Axiomatising KFJ

This subsection emulates Feferman's (1991) axiomatic theory of truth KF to produce a similar type-free axiomatic theory of truth. This theory of truth is very similar to KF, but has different axioms for sentences which are negated conjunctions and negated disjunctions, and this results in interesting novel behaviour. These alternate axioms are motivated by the logic ML_3 and the behaviour of its connectives. In this subsection I shall present and discuss this theory in detail, as well as its relation to KF, and then in the next section I will discuss its relation to ATT.

We start with a language $\mathscr{L} \supseteq \mathscr{L}_A$ and expand this language to $\mathscr{L}_T = \mathscr{L} \cup \{T\}$ where T is intended to be our (type-free) truth predicate for $\mathscr{L}_T$. Again, we work over a base theory B which is sufficiently strong to interpret PA. We take this as implicit within the background, unless otherwise specified, and the resulting

theorems will assume that we have B at our disposable. T is a predicate of Gödel code of formulas, but here we will often drop reference to coding notation for readability. Context should make it clear when we refer to the code of a formula, and when we refer to the formula itself. Let Tr_{At} be the partial (definable in B) truth predicate for $\mathscr{L}$-atomic sentences. We also have the definable predicates $Sent(x)$ and $At_{\mathscr{L}}(x)$ which express "x is the Gödel code of an $\mathscr{L}_T$ sentence" and "x is the Gödel code of an atomic $\mathscr{L}$ formula" respectively. Finally, we again adopt the dot notation specified in Section 1.2.

We now define our new axiomatic theory of truth KFJ with the following axioms:

(KFJ1): $\forall\sigma[At_{\mathscr{L}}(\sigma) \rightarrow (T(\sigma) \leftrightarrow Tr_{At}(\sigma))]$

(KFJ2): $\forall\alpha\forall\beta[Sent(\alpha\dot{\wedge}\beta) \rightarrow (T(\alpha\dot{\wedge}\beta) \leftrightarrow T(\alpha) \wedge T(\beta))]$

(KFJ3): $\forall\alpha\forall\beta[Sent(\alpha\dot{\wedge}\beta) \rightarrow (T(\dot{\neg}(\alpha\dot{\wedge}\beta)) \leftrightarrow$
$\leftrightarrow [T(\dot{\neg}\alpha) \vee T(\dot{\neg}\beta) \vee (\neg T(\alpha) \wedge T(\beta)) \vee (T(\alpha) \wedge \neg T(\beta))])]$

(KFJ4): $\forall\alpha\forall\beta[Sent(\alpha\dot{\vee}\beta) \rightarrow (T(\alpha\dot{\vee}\beta) \leftrightarrow T(\alpha) \vee T(\beta))]$

(KFJ5): $\forall\alpha\forall\beta[Sent(\alpha\dot{\vee}\beta) \rightarrow (T(\dot{\neg}(\alpha\dot{\vee}\beta)) \leftrightarrow$
$\leftrightarrow [(T(\dot{\neg}\alpha) \wedge \neg T(\beta)) \vee (\neg T(\alpha) \wedge T(\dot{\neg}\beta)) \vee (T(\dot{\neg}\alpha) \wedge T(\dot{\neg}\beta))])]$

(KFJ6): $\forall\sigma[Sent(\sigma) \rightarrow (T(\dot{\neg}\dot{\neg}\sigma) \leftrightarrow T(\sigma))]$

(KFJ7): $\forall\sigma[Sent(\dot{\forall}x\sigma(x,\bar{y})) \rightarrow (T(\dot{\forall}x\sigma(x,\bar{y})) \leftrightarrow \forall aT(\sigma(\dot{a},\bar{y})))]$

(KFJ8): $\forall\sigma[Sent(\dot{\forall}x\sigma(x,\bar{y})) \rightarrow (T(\dot{\neg}\dot{\forall}x\sigma(x,\bar{y})) \leftrightarrow \exists aT(\sigma(\dot{a},\bar{y})))]$

(KFJ9): $\forall\sigma[Sent(\dot{\exists}x\sigma(x,\bar{y})) \rightarrow (T(\dot{\exists}x\sigma(x,\bar{y})) \leftrightarrow \exists aT(\sigma(\dot{a},\bar{y})))]$

(KFJ10): $\forall\sigma[Sent(\dot{\exists}x\sigma(x,\bar{y})) \rightarrow (T(\dot{\neg}\dot{\exists}x\sigma(x,\bar{y})) \leftrightarrow \forall aT(\sigma(\dot{a},\bar{y})))]$

(KFJ11): $\forall\sigma[T(\dot{T}(\sigma)) \leftrightarrow T(\sigma)]$

(KFJ12): $\forall\sigma[T(\dot{\neg}\dot{T}(\sigma)) \leftrightarrow T(\dot{\neg}\sigma)]$

(KFJ13): An induction axiom for every formula in the language $\mathscr{L}_T$.

KFJ is an axiomatic type-free theory of truth which is based upon the ML_3 logic specified in Section 5.3.1. We note that the majority of these axioms are identical to the axioms of KF, with the difference being KFJ3 and KFJ5. Just as KF's axioms are based upon the behaviour of the Strong-Kleene connectives, KFJ's axioms are based upon the behaviour of the ML_3 connectives. These connectives differ radically for negated conjunctions and negated disjunctions.

We will prove that KFJ is consistent in Corollary 5.3.2.4, rather than by providing a model for it. Finding models of KFJ is an open question and one that deserves more research – particularly investigation into whether there are fixed-point style constructions for this theory in the same way that KF is based on Kripke's fixed point models.

Question 5.3.2.1. *How can we construct, as a Kripke-style fixed point or otherwise, models $\mathscr{M}$ such that $\mathscr{M} \vDash$ KFJ?*

We will start by investigating the properties of KF. We note that whilst KFJ and KF are extremely similar, the axioms for negated disjunctions and negated conjunctions are where they differ, and this gives rise to a surprising amount of distinction between the theories. This follows from the difference between the Strong-Kleene logic and the logic of ML_3 given in the previous section.

KF is often considered with (exclusively) one of two further axioms: TCons which states $\forall\varphi[\neg(T(\varphi)\wedge T(\dot{\neg}\varphi))]$ and TComp which states $\forall\varphi[Sent(\varphi) \rightarrow (T(\varphi)\vee T(\dot{\neg}\varphi))]$. Informally, these state that the truth predicate is internally consistent or complete, respectively, and notoriously only one of these can be added to KF. The theory of KF + TComp + TCons is inconsistent, since it proves both the Liar sentence and its negation.

We now see what the effect of these axioms on KFJ is. One of the main results is that KFJ + TComp is the same theory (has exactly the same $\mathscr{L}_T$-consequences) as KF + TComp, which proves that KFJ is consistent.

Theorem 5.3.2.2. KFJ + TComp = KF + TComp

Proof. To prove this we show that both of these theories prove the axioms of each other. Since most of the axioms of KFJ and KF are identical, we need only prove this for the axioms governing negated conjunctions and negated disjunctions. We prove the case for negated conjunctions here, and the case for negated disjunctions is similar.

First we work within KFJ + TComp and fix (codes of) formulas φ and ψ. We prove that KFJ + TComp $\vdash T(\dot{\neg}(\varphi \dot{\wedge} \psi)) \leftrightarrow (T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi))$. The axiom KFJ3 tells us:

$$T(\dot{\neg}(\varphi \dot{\wedge} \psi)) \leftrightarrow [T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi) \vee (\neg T(\varphi) \wedge T(\psi)) \vee (T(\varphi) \wedge \neg T(\psi))]$$

We now use the fact that TComp and $\neg T(\sigma)$ implies $T(\dot{\neg}\sigma)$, along with $\wedge$E from classical logic, to deduce:

$$T(\dot{\neg}(\varphi \dot{\wedge} \psi)) \rightarrow [(T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi)) \vee (T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi))]$$

Thus we conclude that KFJ + TComp $\vdash T(\dot{\neg}(\varphi \dot{\wedge} \psi)) \rightarrow (T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi))$. For the converse direction we can simply use KFJ5.

We now work within KF + TComp, again fixing (codes of) fomulas φ and ψ to show that it proves the axiom KFJ3. We use the same fact as above that TComp implies $\neg T(\sigma) \rightarrow T(\dot{\neg}\sigma)$ to deduce:

$$\begin{gathered}[T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi) \vee (\neg T(\varphi) \wedge T(\psi)) \vee (T(\varphi) \wedge \neg T(\psi))] \rightarrow \\ \rightarrow [(T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi)) \vee (T(\dot{\neg}\varphi) \wedge T(\psi)) \vee (T(\varphi) \wedge T(\dot{\neg}\psi))]\end{gathered}$$

The consequent of this formula collapses down to $(T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi))$ using $\wedge$E again, and thus we finally use KF's axiom for negated conjunctions to deduce:

$$[T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi) \vee (\neg T(\varphi) \wedge T(\psi)) \vee (T(\varphi) \wedge \neg T(\psi))] \rightarrow T(\dot{\neg}(\varphi \dot{\wedge} \psi))$$

For the converse direction observe that $T(\dot{\neg}(\varphi \dot{\wedge} \psi) \rightarrow (T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi))$ and thus we can simply use $\vee$I to deduce:

$$T(\dot{\neg}(\varphi \dot{\wedge} \psi)) \rightarrow [T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi) \vee ((\neg T(\varphi) \wedge T(\psi)) \vee (T(\varphi) \wedge \neg T(\psi)))]$$

□

This result immediately entails the following important corollaries about the theory KFJ + TComp, which follow from the literature on KF + TComp:

Corollary 5.3.2.3. KFJ + TComp $\nvdash \bot$ (Cantini, 1989, Theorem 4.3)

Firstly we have an easy proof of the consistency of KFJ + TComp. This importantly entails that KFJ is consistent, since it is a sub-theory of KFJ + TComp.

Corollary 5.3.2.4. KFJ $\nvdash \bot$

We are also able to prove DeMorgan equivalences for KFJ + TComp since these hold within KF + TComp.

Corollary 5.3.2.5.

1. KFJ + TComp $\vdash \forall\varphi\forall\psi[Sent(\varphi) \wedge Sent(\psi) \rightarrow$
$\rightarrow (T(\dot{\neg}\varphi) \vee T(\dot{\neg}\psi)) \leftrightarrow T(\dot{\neg}(\varphi\dot{\wedge}\psi))]$

2. KFJ + TComp $\vdash \forall\varphi\forall\psi[Sent(\varphi) \wedge Sent(\psi) \rightarrow$
$\rightarrow (T(\dot{\neg}\varphi) \wedge T(\dot{\neg}\psi)) \leftrightarrow T(\dot{\neg}(\varphi\dot{\vee}\psi))]$

We also gain information on the arithmetic strength of both PA + KFJ and PA + KFJ + TComp.

Corollary 5.3.2.6. *The arithmetic strength of* PA + KFJ + TComp *is the arithmetic strength of* PA + KF + TComp*, which is the arithmetic strength of* $\mathrm{RA}_{<\epsilon_0}$ (Cantini, 1989, Theorem 9.15).

Corollary 5.3.2.7. *The arithmetic strength of* PA + KFJ *is bounded above by* $\mathrm{RA}_{<\epsilon_0}$.

It seems reasonable to conjecture that the arithmetic strength of PA + KFJ is exactly $\mathrm{RA}_{<\epsilon_0}$ given its similarity to KF, but this is an open question currently.

Conjecture 5.3.2.8. *The arithmetic strength of* PA + KFJ *is exactly* $\mathrm{RA}_{<\epsilon_0}$.

Because the theory KFJ + TComp is the same theory as KF + TComp we set it aside from further discussion.

Instead we now consider the alternative natural addition to KFJ: KFJ + TCons which is the theory KFJ + $\forall\varphi[\neg(T(\varphi) \wedge T(\dot{\neg}\varphi))]$. Unfortunately, the theory KFJ + TCons is inconsistent, since there will always be sentences in KFJ which are both true and their negation true.[8]

Lemma 5.3.2.9. KFJ $\vdash T(\gamma) \wedge T(\neg\gamma)$ *where we define* γ *as the sentence:*
$\gamma \leftrightarrow \neg(1{=}1 \wedge T(\gamma))$.

[8]My thanks and acknowledgements to my examiners for this observation.

Proof. Consider KFJ and the sentence $\gamma \leftrightarrow \neg(1{=}1 \wedge T(\gamma))$ which can be formulated using Gödel's diagonal lemma. First suppose KFJ $\vdash \gamma$. Hence, by logic KFJ $\vdash \neg(1{=}1) \vee \neg T(\gamma)$ and thus by ordinary arithmetic KFJ $\vdash \neg T(\gamma)$. By KFJ11 we hence have KFJ $\vdash \neg T(T\gamma)$. We now use KFJ3 to deduce KFJ $\vdash T(\neg(1{=}1 \wedge T\gamma))$ since KFJ $\vdash T(1{=}1)$. Therefore, by the definition of γ, KFJ $\vdash T(\gamma)$ which is a contradiction.

We therefore know that KFJ $\vdash \neg\gamma$. By logic, hence, KFJ $\vdash 1{=}1 \wedge T(\gamma)$. We know by the definition of γ that thus KFJ $\vdash T(\neg(1{=}1 \wedge T(\gamma)))$. Now, using KFJ3 we deduce KFJ $\vdash T(\neg(1{=}1)) \vee T(\neg\gamma) \vee (\neg T(\gamma) \wedge T(1{=}1)) \vee (\neg T(1{=}1) \wedge T(\gamma))$. By consistency, we cannot have the first, third or fourth disjunct and hence KFJ $\vdash T(\gamma) \wedge T(\neg\gamma)$. □

This lemma can perhaps be seen as an undesirable result for KFJ, since it shows there are sentences which the theory refutes, but proves are both true and their negation true. We can use this to prove that KFJ + TCons is inconsistent.

Corollary 5.3.2.10. KFJ + TCons $\vdash \bot$.

Proof. KFJ + TCons $\vdash \forall\varphi[\neg(T(\varphi) \wedge T(\dot{\neg}\varphi))]$, but this contradicts Lemma 5.3.2.9 above. □

This is the final result of this section, and highlights some interesting, albeit perhaps unwanted, differences between the behaviour of KFJ and KF. In the next section we look at the connection between ATT and KFJ and what this suggests about the connection between typed and type-free truth more generally.

5.4 ATT and KFJ

We have now seen many details about the axiomatic type-free theory of truth KFJ. I introduced this theory by a three-valued logic ML_3 and showed that we can use this to build Kripke-style fixed points. The theory was motivated, however, by our theory of Axiomatic Typed Truth from Section 5.2. In this Section we investigate the connection between the two theories of truth. We will see that KFJ is something of a 'type-free' variant of ATT and that if a sentence is true according to ATT then it is true for KFJ. Further, if KFJ thinks a non-zero rank formula is true, then ATT believes that it is true at some level as well.

We will be slightly informal in our translations between the language $\mathscr{L}_{Tr}$ of ATT and $\mathscr{L}_T$ of KFJ, and often use the same formula (meta)variable φ interchangeably between both languages, when strictly we mean $\varphi_{\mathscr{L}_{Tr}}$ when φ is an $\mathscr{L}_{Tr}$-formula and $\varphi_{\mathscr{L}_T}$ when φ is an $\mathscr{L}_T$-formula.

Our translation is defined inductively on the complexity of φ. If $\varphi_{\mathscr{L}_{Tr}}$ is an $\mathscr{L}_{Tr}$-formula, then we translate the $\mathscr{L}$ symbols in φ directly across as their equivalent symbols in $\mathscr{L}_T$. If $\varphi_{\mathscr{L}_{Tr}}$ is a formula of the form $R(\ulcorner\psi_{\mathscr{L}_{Tr}}\urcorner, x)$ then $\varphi_{\mathscr{L}_T}$ is the formula:

$$[x \mathring{>} \mathring{0} \rightarrow (T(\ulcorner\psi_{\mathscr{L}_T}\urcorner) \vee T(\ulcorner\neg\psi_{\mathscr{L}_T}\urcorner))] \wedge [x \mathring{=} \mathring{0} \rightarrow \neg(T(\ulcorner\psi_{\mathscr{L}_T}\urcorner) \vee T(\ulcorner\neg\psi_{\mathscr{L}_T}\urcorner))]$$

If $\varphi_{\mathscr{L}_{Tr}}$ is a formula of the form $Tr(\ulcorner\psi_{\mathscr{L}_{Tr}}\urcorner, x)$, then $\varphi_{\mathscr{L}_T}$ is the formula $x \mathring{>} \mathring{0} \rightarrow T(\ulcorner\psi_{\mathscr{L}_T}\urcorner)$. Similarly, given $\varphi_{\mathscr{L}_T}$ in $\mathscr{L}_T$ we will translate all $\mathscr{L}$-symbols as their equivalent in $\mathscr{L}_{Tr}$. If φ is of the form $T(\ulcorner\psi_{\mathscr{L}_T}\urcorner)$, then $\varphi_{\mathscr{L}_{Tr}}$ is the formula $Tr(\ulcorner\psi_{\mathscr{L}_{Tr}}\urcorner, n)$ where n is such that $\mathrm{B} + \mathrm{ATT} \vdash R(\ulcorner\psi_{\mathscr{L}_{Tr}}\urcorner, n)$.

We now show that these two theories believe the same formulas are true, for the positive rank formulas. We will state and prove this formula slightly informally, without reference to coding, for ease of reading.

Theorem 5.4.1. *If φ is a formula such that $\mathrm{B} + \mathrm{ATT} \vdash R(\varphi, n) \wedge n \mathring{>} \mathring{0}$, then $\mathrm{B} + \mathrm{ATT} \vdash \mathrm{Tr}(\varphi, n)$ if and only if $\mathrm{B} + \mathrm{KFJ} \vdash T(\varphi)$.*

Proof. We prove this by induction on the complexity of φ. In particular, our induction hypothesis is that the theorem holds for all ψ of lower complexity, and all subformulae of such ψ and the negation of these subformulae of ψ.

- If φ is an atomic truth-free formula, then $\mathrm{B} + \mathrm{KFJ} \vdash T(\varphi) \leftrightarrow Tr_{At}(\varphi)$ by KFJ1. We also know by ATT1 that $\mathrm{B} + \mathrm{ATT} \vdash Tr_{At}(\varphi) \leftrightarrow Tr(\varphi, 1)$ and hence we are done.

- If φ is of the form $\alpha \wedge \beta$, then we know by KFJ2 that $\mathrm{B} + \mathrm{KFJ} \vdash Tr(\varphi) \leftrightarrow (Tr(\alpha) \wedge Tr(\beta))$. By induction we have that $\mathrm{B} + \mathrm{KFJ} \vdash Tr(\alpha) \wedge Tr(\beta)$ if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(\alpha, n) \wedge Tr(\beta, n)$ (using ATT8 if necessary) and this holds by ATT2 if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(\varphi, n)$.

- The case for φ of the form $\alpha \vee \beta$ is similar and hence omitted here.

- If φ is of the form $\neg(\alpha \wedge \beta)$ such that $\mathrm{B} + \mathrm{ATT} \vdash R(\neg(\alpha \wedge \beta), n)$ then we know by the axioms of rank that $R(\alpha, n)$ or $R(\beta, n)$.

We first assume that $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg(\alpha \wedge \beta), n)$, so $n \mathring{>} \mathring{0}$ and $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\alpha, n) \vee \neg Tr(\beta, n)$. If $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\alpha, n)$ and $R(\alpha, n)$, then $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg\alpha, n)$ and by inductive assumption $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg\alpha)$. Hence by KFJ3 $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg(\alpha \wedge \beta))$ and we are done. Similarly if $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\beta, n)$ and $R(\beta, n)$ we are finished.

Hence we are left with the case where $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\alpha, n)$ and $R(\beta, n)$ (the other case is symmetrical). By Corollary 5.2.4.12 $\mathrm{B} + \mathrm{ATT} \vdash Tr(\beta, n) \vee Tr(\neg\beta, n)$. If $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg\beta, n)$ then we know by the argument above that the statement is proven, so we assume $\mathrm{B} + \mathrm{ATT} \vdash Tr(\beta, n)$. We deduce $\mathrm{B} + \mathrm{ATT} \nvdash Tr(\alpha, n)$ by consistency and hence by inductive assumption $\mathrm{B} + \mathrm{KFJ} \nvdash T(\alpha)$ and $\mathrm{B} + \mathrm{KFJ} \vdash T(\beta)$. Thus we have $\mathrm{B} + \mathrm{KFJ} \vdash \neg T(\alpha) \wedge T(\beta)$ and by KFJ3 $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg(\alpha \wedge \beta))$.

We now assume $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg(\alpha \wedge \beta))$. We thus know from KFJ3 that $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg(\alpha \wedge \beta)) \leftrightarrow [T(\neg\alpha) \vee T(\neg\beta) \vee (\neg T(\alpha) \wedge T(\beta)) \vee (\neg T(\beta) \wedge T(\alpha))]$. First we assume $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg\alpha)$. In this case we know by induction that entails $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg\alpha, n)$ and hence $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\alpha, n)$. Thus we know $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr((\alpha \wedge \beta), n)$ and hence $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg(\alpha \wedge \beta), n)$ since $R(\neg(\alpha \wedge \beta), n)$. The case for $\mathrm{B} + \mathrm{KFJ} \vdash Tr(\neg\beta)$ is identical.

We finally consider the case where $\mathrm{B} + \mathrm{KFJ} \vdash \neg Tr(\alpha) \wedge Tr(\beta)$ (since the only remaining case is symmetrical) with the additional assumption that $\mathrm{B} + \mathrm{KFJ} \nvdash T(\neg\alpha), T(\neg\beta)$. We hence know $\mathrm{B} + \mathrm{ATT} \vdash Tr(\beta, n)$ and $\mathrm{B} + \mathrm{ATT} \nvdash Tr(\neg\alpha, n), Tr(\neg\beta, n)$ by inductive hypothesis. We thus conclude $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\neg\alpha, n)$. If $R(\alpha, k)$ where $k \mathring{>} \mathring{0}$, then $\mathrm{B} + \mathrm{ATT} \vdash Tr(\alpha, n)$ and hence by inductive hypothesis $\mathrm{B} + \mathrm{KFJ} \vdash T(\alpha)$ which is contradiction. Therefore $R(\alpha, \mathring{0})$. Hence $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\alpha, n)$ and so $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\alpha \wedge \beta), n)$. Now, since $R(\neg(\alpha \wedge \beta, n))$ we have that $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg(\alpha \wedge \beta), n)$.

- If φ is of the form $\neg(\alpha \vee \beta)$ such that $R(\neg(\alpha \vee \beta), n)$, then again we know $R(\alpha, n)$ or $R(\beta, n)$.

  We first assume that $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg(\alpha \vee \beta), n)$ and thus $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\alpha, n) \wedge \neg Tr(\beta, n)$. Assuming without loss of generality that $R(\alpha, n)$ we deduce that $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg\alpha, n)$ and $\mathrm{B} + \mathrm{ATT} \nvdash Tr(\beta, n)$. Hence by inductive hypothesis we know that $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg\alpha)$ and $\mathrm{B} + \mathrm{KFJ} \nvdash T(\beta)$. Therefore $\mathrm{B} + \mathrm{KFJ} \vdash \neg T(\beta)$ and hence by KFJ5 $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg(\alpha \vee \beta)$.

We now assume $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg(\alpha \wedge \beta))$. We thus deduce with KFJ5 that $\mathrm{B} + \mathrm{KFJ} \vdash (T(\neg\alpha) \wedge \neg T(\beta)) \vee (\neg T(\alpha) \wedge T(\neg\beta)) \vee (T(\neg\alpha) \wedge T(\neg\beta))$. If $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg\alpha) \wedge T(\neg\beta)$ then by inductive hypothesis we know $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg\alpha, n) \wedge Tr(\neg\beta, n)$ and hence $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg(\alpha \vee \beta), n)$.

We thus assume $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg\alpha) \wedge \neg Tr(\beta)$ (the proof for the remaining case is symmetric). We hence derive $\mathrm{B} + \mathrm{KFJ} \nvdash T(\beta)$ and by inductive hypothesis $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg\alpha, n)$ and $\mathrm{B} + \mathrm{ATT} \nvdash Tr(\beta, n)$. Therefore $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\alpha, n)$ and $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\beta, n)$ and hence $\mathrm{B} + \mathrm{ATT} \vdash Tr(\neg(\alpha \vee \beta), n)$.

- If φ is of the form $\forall x \psi(x)$, then we know by KFJ7 that $\mathrm{B} + \mathrm{KFJ} \vdash T(\varphi) \leftrightarrow (\forall a T(\psi(\dot{a}))$. By induction we have that $\mathrm{B} + \mathrm{KFJ} \vdash T(\psi(\dot{a}))$ for each a if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(\psi(\dot{a}), n)$ for each a (using ATT7 if necessary). This holds if and only if $\mathrm{B} + \mathrm{ATT} \vdash \forall a\, Tr(\psi(\dot{a}), n)$ and hence by ATT5 this holds if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(\varphi)$.

- The case for φ of the form $\exists x \psi(x)$ is similar and hence omitted here.

- If φ is of the form $\neg\forall x \psi(x)$ then we know by KFJ8 and KFJ9 that $\mathrm{B} + \mathrm{KFJ} \vdash T(\varphi)$ if and only if $\mathrm{B} + \mathrm{KFJ} \vdash T(\exists x \psi(x))$, and thus by the previous case this holds if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(\exists x \psi(x), n)$. We know by Lemma 5.2.4.4 that $\mathrm{B} + \mathrm{ATT} \vdash Tr(\exists x \psi(x), n)$ if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(\varphi, n)$.

- The case for φ of the form $\neg\exists x \psi(x)$ is similar and hence omitted here.

- If φ is of the form $Tr(\psi)$, then we know that $\mathrm{B} + \mathrm{KFJ} \vdash T(T(\psi))$ if and only if $\mathrm{B} + \mathrm{KFJ} \vdash T(\psi)$ and by inductive hypothesis this holds if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(\psi, n-1)$. By T7 this holds if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(Tr(\psi, n-1), n)$ and hence we are done.

- Finally if φ is of the form $\neg Tr(\psi)$, then we know $\mathrm{B} + \mathrm{KFJ} \vdash T(\varphi)$ if and only if $\mathrm{B} + \mathrm{KFJ} \vdash T(\neg\psi)$. This holds by our inductive hypothesis if and only if $\mathrm{B} + \mathrm{ATT} \vdash T(\neg\psi, n-1)$ and hence $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\psi, n-1)$. We now use Theorem 5.2.4.2 to deduce that this holds if and only if $\mathrm{B} + \mathrm{ATT} \vdash Tr(\varphi, n)$.

□

Corollary 5.4.2. *For any $\mathscr{L}_{\mathrm{Tr}}$-formula φ, if* $\mathrm{B} + \mathrm{ATT} \vdash \exists x \mathrm{Tr}(\varphi, x)$*, then* $\mathrm{B} + \mathrm{KFJ} \vdash T(\varphi_{\mathscr{L}_T})$.

Proof. If $\mathrm{B} + \mathrm{ATT} \vdash \exists x\, Tr(\varphi, x)$, then $\mathrm{B} + \mathrm{ATT} \vdash Tr(\varphi, n) \wedge R(\varphi, n)$ for some n. The corollary then follows from Theorem 5.4.1. □

The theorem and corollary above show that KFJ is something of a type-free variant of ATT. For the non-zero rank formulas, ATT and KFJ believe that the same formulas are true, but KFJ makes no distinction on levels of the truth predicate. This comes with some natural benefits for KFJ: a more natural 'type-free' theory of truth for one thing and an interesting background logic that provides more classical valuations than usual three-valued logics. We did not add a rank predicate to our language for KFJ, as it has no need to assign levels to sentences, but Theorem 5.4.1 shows that all the nice behaviour of ATT for the positive-rank formulas holds within KFJ as well. This means a T-Schema, internal completeness, and internal consistency hold within KFJ for all the positive-rank formulas.

This is an example of deep similarities between typed and type-free theories of truth, which can be seen in many formal results. For example, Feferman (1991) shows that both $\mathrm{RT}_{<\epsilon_0}$ and KF have the same arithmetical strength and Halbach (1997) shows that a transfinite Tarskian theory of truth can be embedded into Kripke's least fixed point. We can view Theorem 5.4.1 as showing a proof-theoretic connection between ATT and KFJ; a further example that the typed and type-free approaches to truth are not so disconnected as it might first appear.

We can ask whether this connection between the two theories extends further, might it be the case that ATT is actually interpretable within KFJ?

Question 5.4.3. *Is* ATT *interpretable within* KFJ*?*

One key difference between the two theories is that KFJ can also view some of the zero-rank formulas as true as well, whereas ATT views all of these as untrue. An example of this is the sentence σ: $T(\ulcorner \forall x T^x(0 = 0) \urcorner)$ where T^x informally denotes x-many occurrences of the truth predicate.[9] We know $B + \mathrm{KFJ} \vdash \sigma$ since $\mathrm{KFJ} \vdash \forall a T(\ulcorner T^{\dot{a}}(0 = 0) \urcorner)$, but $\sigma_{\mathscr{L}_{Tr}}$ has rank 0 according to B + ATT. This is because $Tr^a(0 = 0, a)$ has rank $a + 1$ for each a, and hence $\forall x\, Tr^x(0 = 0, x)$ has rank of the supremum of $\{a + 1 : a \in M\}$ which by our agreement is 0, since this set is unbounded. Therefore $\mathrm{B} + \mathrm{ATT} \vdash \neg Tr(\ulcorner \forall x\, Tr^x(0 = 0, x) \urcorner, n)$ for any n.

[9] More formally, we can introduce a binary primitive recursive function f such that $f(n, \varphi) = T^n(\ulcorner \varphi \urcorner)$. Halbach (2011, p. 157-8) provides further details on this function and this formalisation.

This gain in alethic provability for KFJ is offset by areas where it appears too much, such as Lemma 5.3.2.9. KFJ refutes a sentence, but also proves that it is true, which appears highly questionable of a theory of truth and akin to a paradox, particularly when it also proves this sentence is false. This is not technically inconsistent, but does appear deeply unintuitive behaviour for a theory of truth.

Given such remarks, and the motivation of this chapter, I am inclined to see ATT as a formally adequate theory of truth. This theory features all the desirable properties of a truth predicate for the sentences of the base language and all 'reasonable' sentences built from these with the truth predicate. Such properties are an extended T-Schema from Chapter 2, compositionality and completeness. The theory retains consistency due to its typed approach, which can be defended philosophically be appeal to a contextual notion of truth. The syntactic shortcomings of a usual typed approach are overcome here and the notion of 'truth-aptness' defends ATT's behaviour with respect to rank 0 sentences – those already troubling philosophical sentences such as paradoxes and those which quantify absolutely. Further, we have an appealing model-theoretic interpretation of ATT with a model of Tarski's semantic truth hierarchy. Whilst KFJ also boasts some attractive features,the result of Lemma 5.3.2.9 means that it loses out for formal adequacy under my understanding.

5.5 Conclusion

In this chapter I have presented two new axiomatic theories of truth: ATT and KFJ. I have shown that ATT is a broadly contextual theory of truth, in the Tarskian tradition, where the truth predicate is treated as a binary relation between sentences and levels. This has the intended interpretation that $Tr(\sigma, n)$ if and only if σ is true at level n. The theory comes packaged with a ranking of sentences and the theory has many desirable properties for sentences with non-zero rank such as the T-Schema, compositionality, internal completeness and internal consistency. The rank 0 sentences, I argue, should be interpreted as 'not truth-apt' and these sentences are either paradoxical or quantify over absolutely all levels of the truth predicate. This prompts discussion on the adequacy of such treatment, as well as its philosophical applications.

I have then introduced a new three-valued logic ML_3, based on the behaviour of ATT's truth predicate, which maximises classical valuations as much as possible.

This logic can be used to generate Kripke-style fixed points, although interestingly this procedure is not monotone. This theory can be axiomatised to produce KFJ, which is a KF-like theory of truth. It is interesting to note that a small change in the axioms of KF results in some quite different behaviour in some ways, and very similar behaviour in others. KFJ is a type-free theory of truth, which is closely connected to ATT, and can speak about the truth of sentences which quantify over absolutely all iterations of the truth predicate. The downside of this approach is that the models of this theory behave questionably, and we no longer have a theory which deals with the semantic paradoxes.

I conclude that these remarks show that ATT is an adequate axiomatic theory, for formal purposes, in the sense of Chapter 4. This supports deflationism about truth and partially answers the question at the end of Chapter 4.

I have proven many results about the two theories, particularly their alethic properties, but many open questions resulting from this research remain. I have detailed many of these formal questions and conjectures throughout the chapter, and have provided a summary of these below, but I believe that there are also many interesting philosophical questions remaining from this research. In particular, it is interesting to explore how much ATT relates to more traditional contextual theories of truth, and to what extent its approach to paradox is immune to the 'revenge' paradox. Whilst the theory is formally consistent, it does advocate that the liar sentence is provable, and hence not true, which are usual grounds for deriving that the sentence therefore must be true, and we are back in paradoxical territory. Another question of interest is the relation between provability and truth in ATT. There are sentences which are provable, but not true in the theory (many rank 0 sentences) and it seems worthy of comment how notions of proof and truth fall apart for ATT. A philosophical interpretation and defence of this would be of interest.

The behaviour of KFJ is also interesting, and leaves open the question of whether we can formulate a theory of truth 'partial KFJ' (PKFJ) which stands in the same relation to KFJ as PKF does to KF. It would be interesting to see how such a development would go, and the extent to which it would differ to PKF.

Question 5.5.1. *Is there a* 'PKFJ' *variant of* KFJ*, which is analogous to the* PKF *variant of* KF*?*

There are also questions over what other three-valued logics can be used to

motivate KF-like theories of truth, and how these behave in relation to KF and KFJ. Bolc and Borowik (1992) detail a great number of three-valued logics beyond the standard Strong and Weak Kleene schemes, and it would be of interest to see whether these are suitable for producing new truth theories too, particularly those with alternative rules for negation.

I leave these questions as open for further research, for they go beyond my aim in this thesis. In this chapter I have provided an axiomatic theory of truth, ATT, that according to Chapter 4 is a deflationary theory of truth. I hope that I have shown that this theory can be seen as formally adequate, in the sense of Chapter 4. This leaves open the question of whether such a theory is philosophically adequate, however. By philosophical adequacy, I mean a theory of truth which is suitable for philosophical purposes. My aim in the next chapter is to show that even a very simple deflationary theory of truth is suitable philosophically, since it can accommodate the explanatory power of a pluralist theory of truth. This answers the question at the end of Chapter 4 positively: a deflationary theory can be formally and philosophically adequate.

Open Questions and Conjectures

5.2.2.2 What is the recursive complexity of the rank relation R(x,y)?

5.2.4.8 Is the arithmetic strength of PA + ATT $RA_{<\epsilon_0}$?

5.2.4.10 Is PA + ATT^{-} proof-theoretically conservative over PA?

5.2.5.3 Can we formulate Yablo-Visser style paradoxes in $\mathscr{L}_{Tr}$, and if so, how does ATT deal with these?

5.3.1.1 Are there logics analogous to ML_3 which have four or more possible semantic values?

5.3.2.1 How can we construct, as a Kripke-style fixed point or otherwise, models $\mathscr{M}$ such that $\mathscr{M} \vDash$ KFJ?

5.3.2.8 Is the arithmetic strength of PA + KFJ exactly $RA_{<\epsilon_0}$?

5.4.3 Is ATT interpretable within KFJ?

5.5.1 Is there a 'PKFJ' variant of KFJ, which is analogous the PKF variant of KF?

Chapter 6

Deflating Alethic Pluralism

In Chapter 4 I concluded with the question of whether we can have a deflationary theory that is formally and philosophically adequate. In the previous chapter I have aimed to show that we have an axiomatic (and thus deflationary) theory of truth which is formally adequate. In this chapter I shall aim to show that even a very simple deflationary theory of truth can be philosophically adequate. I will do this by comparison to pluralist theories of truth, and will argue that a variant of deflationism can accommodate all the philosophical uses of a plural theory of truth. Since these theories give truth great explanatory power, and are philosophically adequate, I claim that deflationary theories of truth can be as well. This provides an answer to Chapter 4, half of which is provided by Chapter 5, that a deflationary theory can be formally and philosophically adequate. I will take these results to provide an argument, to be given in Chapter 7, for deflationism about truth, using formal truth theory.

Chapter Abstract

I present a deflated understanding of pluralim about truth: a theory which combines the metaphysical simplicity of a deflationary account of truth with the explanatory power of a plural account of truth. This theory endorses a single deflationary truth property, but admits there can be many truth-like properties: properties extensionally equivalent with the truth property for fragments of the language. I argue that this theory avoids traditional worries with alethic pluralism since it faithfully captures our unary use of the truth predicate in everyday language, but also enables the deflationist to explain key features of particular domains of discourse with the truth predicate. If pluralism is motivated by the apparent explanatory

role truth plays in particular domains, then I claim that deflationary alethic pluralism is a compelling position for the would-be pluralist.

6.1 Introduction

There has been a significant recent development of plural approaches to truth: theories of truth which admit that there are a plurality of ways in which sentences can be true. This is in comparison to monist theories of truth which typically analyse the concept of truth with only a single property of truth. Advocates such as (Crispin) Wright (1998), Lynch (2009), Pederson and (Cory) Wright (2013) and Edwards (2011) observe the general failures of traditional monist theories of truth and diagnose that monism is at issue. There appears to be much agreement with this point. In *From One to Many: Recent Work on Truth* Wyatt and Lynch (2016) note that:

> The last decade has seen the development of both novel versions of traditional theories of truth as well as several strikingly new kinds of account ... An underappreciated thread running through these views, we'll argue, is a certain pluralizing tendency.

This pluralising tendency captures the intuition that sentences about wildly different subject matters are true in different ways. Alethic pluralism offers a substantive explanation of why this is the case and, moreover, the distinctive features of certain kinds of true sentences. The explanatory benefits of pluralism come with a weighty metaphysical cost, however, particularly in comparison to lightweight deflationary alternatives that have been offered by philosophers such as Horwich (1998), Künne (2003), and Quine (1986).[1] In addition to this, current plural theories of truth face significant challenges accounting for all our uses of 'truth' in language, which I will argue result from our unary usage of the truth predicate. Again, this problem is neatly avoided by a deflationary theory of truth.

Given such remarks, it seems advantageous to accommodate the explanatory powers of a plural attitude to truth in a deflationary account of truth, and this is my aim in this chapter. In Section 6.2 I will introduce plural theories of truth in more detail, and set out the issues facing them. I will argue that these result from

[1] In Chapter 3 I provide more details on these theories and discuss what deflationism about truth is in general.

the unary way we use the truth predicate in language and take this to motivate a deflationary treatment of truth. I will then in Section 6.3 introduce a deflationary alethic pluralism, where the pluralism is not a plurality of truth properties, but of properties extensionally equivalent with truth for particular discourses. In Section 6.3.1 I discuss what a deflationary alethic pluralist can say about these properties and their associated domains of discourse and in Section 6.3.2 show they can fulfil the explanatory role of the plural truth properties whilst avoiding challenges traditionally laid against alethic pluralism. Finally, in Section 6.3.3 I compare how this theory offers a significant advantage to other pluralist theories on offer. I shall conclude that if one is motivated towards a plural account of truth, then there are compelling reasons to adopt my deflationary pluralist variant instead. This shows a deflationary theory of truth can be philosophically adequate, in the sense of Chapter 5.

6.2 Alethic Pluralism

Pluralism about truth is, at its core, the position that truth can have many natures, often proposed with the stance that there are a number of different truth properties. Alethic pluralism has been advocated in various guises by philosophers such as Wright (1998), Lynch (2009), Pedersen (2014), and Edwards (2011). The plural position contests traditional monist accounts of truth, according to which there is only one truth property, and instead claims that there can be different truth properties[2] for different sentences. Sentences are categorised into 'domains of discourse', which can be thought of as different semantic categories, and the pluralist admits that the truth property for one domain of discourse may differ to the truth property for another domain. The pluralist admits a variety of domains of discourse, each of which can utilise a different truth property, and which can behave in different ways.

An example of this is that a sentence about particle physics (P) might belong to a worldly domain of discourse, whereas a sentence about the legality of recreational drugs (R) would belong to a legal domain of discourse. The pluralist can claim that P is true when there is some kind of worldly fact that corresponds to P, that

[2]This is a rough-and-ready formulation of alethic pluralism, and slightly conflicts with some formulations (e.g. Lynch's (2009) functional theory) of pluralism. I discuss Lynch's theory shortly, but the distinction between monism and pluralism in terms of properties is a useful linguistic generalisation and one I shall abuse later.

the truth property for sentences belonging to the worldly domain of discourse is the same as given by a traditional correspondence theory of truth.[3] The pluralist can claim that R is not true when it corresponds to a fact, however, but instead when it coheres with the current body of law, and that the truth property for the legal domain of discourse is the same as given by a coherence theory of truth.[4]

That the pluralist is able to make, and make sense of, such claims appears to put her in a better position than traditional monist theories of truth, which admit only one property of truth. As Lynch (2009) observes, these monist accounts of truth appear to provide adequate analysis for only parts of our language, and face issues when generalised beyond these. A correspondence theory of truth fits intuitions about sentences about physical laws cleanly, where we might have genuine worldly facts, but faces difficulties when extended beyond these. It is tough to claim that legal laws are true in the same way as physical laws, since legal laws vary over times and places, but physical laws are generally held to be universal. It seems far more plausible that a (legal) law is true when it holds a coherent position within the rest of the body of law, but a coherency theory of truth now struggles to provide an adequate truth property for worldly sentences. For instance, Thagard (2007) argues that under a coherence theory of truth worldly sentences cannot represent the world, but merely relate to other representations, contrary to natural intuitions about worldly sentences. This looks like it requires some form of resolution, and the pluralist has an easy response, which is to admit that these different kinds of sentences really are just true in different ways.

An alternative example can be seen by distinguishing between scientific truths and fictional truths. It seems like scientific truths have existential import. If it is true that atoms consist of a nucleus and electrons, then we take it to imply that atoms, nuclei, and electrons exist. On the other hand, fictional truths do not appear to have existential import. If it is true that Frodo owns the sword Sting, then we would not take it to mean that Frodo or Sting exists. Of course, many strategies exist for avoiding such issues, but the pluralist is able to meet such intuitions head-on and explain why truths for different kinds of sentences can behave differently, because *truth* for these sentences can be different. Monist

[3] A correspondence theory of truth tells us that truth consists in a truthbearer corresponding with (a part of) the world - often denoted as a fact. Patterson (2003) provides a more precise discussion of such theories and what they amount to.

[4] A coherence theory of truth is loosely that truth consists in coherence with some already specified set of truthbearers. Walker (2001) provides a discussion and defence of such theories.

theories of truth seem unable to account for these differences, and need to explain them away. The pluralist cites such partial failures of monist theories of truth as evidence for her position.

These claims about the plurality of truth have been specified in different ways, and I would like to take a moment to sketch three such theories, in order to contrast them with my own position later. Wright's (1998) original formation of pluralism is that we have a single concept of truth, but multiple properties of truth which exemplify it. The concept of truth is governed by numerous platitudes about truth, such as the equivalence schema and being the aim of belief, which all the truth properties satisfy. The truth properties also satisfy their own distinct further conditions, relative to the domain of discourse to which they apply. Perhaps the concept of being metallic is a useful example here. We have numerous platitudes for what it is to be a metal, such as conductivity and the ability to form alloys, and many properties which exemplify this, such as being iron, lead, or tin, and these properties also satisfy their own distinct further conditions. Wright's account of truth is not dissimilar, we have a single concept of truth, and many properties of truth (such as correspondence, coherence, and warranted assertibility) which fall under this concept.

Lynch (2009) has an alternative account of pluralism, a functionalist account, analogous to functionalism in philosophy of mind. Functionalism in philosophy of mind proposes that certain properties are functional properties, meaning that the same property can be realised in different ways by different creatures. Pain in humans may be physically realised very differently to pain in octopuses, but both creatures experience a property of pain, since they experience a property which plays the 'pain'-role. For Lynch, truth is similarly a functional property and the property of truth can be realised in different ways within different domains of discourse. If a property plays the 'truth'-role for a particular domain of discourse, then it is the truth property for that domain.

The final variant of alethic pluralism that I would like to consider is Edwards' (2013b) Simple Determination pluralism, which presents an analogy between truth and winning. The aim of every game is different and, depending upon the game being played, the rules for determining the winner of that game differ as well. There is a single property of winning, but players need to satisfy a further property to determine the winner of a particular game. For Edwards there is similarly a single truth property, much like the property of winning, but depending upon the

domain of discourse, sentences need to satisfy a further property to determine whether they are true or not. The rules for determining truth differ for different domains of discourse.

Whilst these theories offer different pluralisms about truth, they all share certain benefits associated with a pluralist position about truth. The pluralist avoids the worry that, due to the wildly different ways that sentences appear to be true, there can be no single account of truth as the traditional theories aim to provide. Wright (2005, p. 4) writes of the motivation for pluralism:

> allegiance to alethic monism is what generates explanatory inadequacy. So, while traditional inflationary approaches successfully explain how individual propositions in certain domains of discourse can be true, those approaches fail to specify *the* nature of truth because they run up against counterexamples when attempting to generalize across all domains.

Of course, monist theories have developed responses to try and sidestep, or bow, to such issues as they come, but the pluralist is in a good position to meet this behaviour at once, and can further explain why and how it occurs. The pluralist is in an excellent position to explain distinctive features of true sentences within certain domains of discourse, such as worldly, moral, aesthetic, mathematical, and even fictional truths: notoriously distinctive areas. There is great motivation for a pluralist position about truth, given its great explanatory resource to philosophers' toolkits.

6.2.1 Problems for Pluralism

The pluralist may boast such attractive features of her view, but she also faces some substantive challenges to her position. I wish to set these out in some detail, in order to argue that their root cause is in the non-plural usage of the truth predicate in everyday language.[5] The first of these problems besetting pluralism is whether it is really possible to characterise every sentence as belonging to a unique domain of discourse. For a sentence to be true or false it needs access to a truth property, and for that it needs to belong to a particular domain of

[5]Of course, this is not an exhaustive list of problems which have been set upon alethic pluralism, but a summary of some of the more pressing universal challenges which affect each of the theories I have sketched above.

discourse. It may be easy to classify '$1 + 1 = 2$' as a mathematical sentence and 'murder is wrong' as a moral sentence, but other atomic sentences appear more problematic. As an example, 'space is non-Euclidean' appears to involve concepts from both mathematical discourse and physical discourse, and hence belong to both of these domains of discourse. An example adapted from Sher (2005), who first formulated this challenge, is the sentence 'causing someone to feel pain is wrong' which involves notions of causality, mental states, and morality. It appears that the pluralist needs to provide some kind of non-arbitrary classification for these problematic sentences, and this is not an easy task.

This problem of sentences seemingly belonging to multiple domains of discourses is exacerbated when compound sentences are considered. Tappolet (2000) concisely sets out that pluralism has issues with accounting for the truth of sentences which are a connection of two clauses, when each clause belongs to a different domain of discourse. Consider the sentence 'murder is wrong and $1 + 1 = 2$.' Broadly speaking, 'murder is wrong' is true for the pluralist because it satisfies a particular moral truth property True_M. Accordingly we write 'murder is wrong' is true_M. The sentence '1+1=2' is true for the pluralist because it satisfies a different mathematical truth-property True_N, '1+1=2' is true_N. The question is which truth-property their conjunction 'murder is wrong and $1+1 = 2$' satisfies. It seems implausible that it would be either True_M or True_N, since neither is adequate for the other conjunct, and thus it appears that the conjunction is true because it satisfies some third truth property True_3. If this is the case, however, then it seems that we can apply the simple rule of logic that A&B is true implies A is true and B is true to find that 'murder is wrong' is true_3 and '$1 + 1 = 2$' is true_3. We can extend this argument to sentences from every domain, and find that each true sentence is true_3. Therefore it appears that we require only one truth-property, true_3, in contrast to the pluralist's starting point.

A similar problem faced by the pluralist about truth is detailed by Wright (2005) and is the struggle alethic pluralism has with accounting for general sentences, which refer to sentences from potentially all domains of discourse. Horwich (1998) argues that one of the main functions of the truth predicate is its ability to form general sentences. These are sentences like: 'everything written in this thesis is true' or: 'every sentence is true or false'. These sentences quantify over sentences from many different domains of discourse and, similarly to the problem of compound sentences, it seems that we need a 'general truth' property true_G

for analysis of these sentences. No single truth property will do, since it will only apply to sentences from its corresponding domain of discourse. If every sentence is true$_G$ or false$_G$, then again it appears that any individual sentence is true$_G$ or false$_G$, and this is sufficient as the pluralist's only truth property.

I do not wish to imply that these problems for the pluralist cannot be overcome. Pluralists have offered defences against these arguments, for instance Lynch and Edwards both argue that their views are unaffected by worries around mixed compound sentences and general sentences. Lynch (2009) proposes that logically complex sentences are "*plainly* true", in the sense that the truth property itself plays the truth-role for these sentences. The truth value of such compound sentences supervenes upon its atomic sentences and thus will still depend upon other properties playing the truth-role for these atomic sentences. Alternatively, Edwards (2011) introduces a notion of 'order of determination', where a conjunction is determined by its instances, and argues that hence a conjunction is determined to be true by the truth properties of its conjuncts, with no further truth property required. Pederson and (Cory) Wright (2013) have adapted (Crispin) Wright's original view to include disjunctive truth properties, which are the union of other truth properties, which fulfil the role required of mixed and general truth properties. Pedersen (2010) follows a Lewisian (1986) view of properties and argues that disjunctive truth properties are abundant properties, as opposed to the sparse basic truth properties. This ensures that the focus for the pluralist is still on the sparse fundamental truth properties, but that abundant truth properties can manifest the concept of truth for logically complex sentences. Finally, in response to the problem of mixed atomic sentences, Wyatt (2013) proposes that pluralists should admit that these sentences belong to more than one domain, whereas Lynch (2005) argues that these examples can be paraphrased away into a single domain of discourse.

Such a litany of responses shows that the pluralist can respond to such challenges, even if not all of these replies can be jointly consistently held, but the route to solving these problems is to introduce additional consequential metaphysical structure. Moreover, such specifications still do not quite show that pluralism's motivations are not compromised by such additions. It appears that a disjunctive truth property of every sparse truth property *à la* Pederson and Wright or a plain truth property *à la* Lynch will be adequate as a monist theory of truth. Pluralists are in a comfortable position to argue that these are not the basic truth property

for their theory, and merely emerge from it, but it shows that a pluralist is led to eventually provide a unary account of truth. The challenges show, whether they can be overcome or not, that there are significant differences between the way we use the truth predicate in ordinary language, and the pluralist's account of multiple truth properties. Our natural language truth predicate can be used freely, without consideration of domains of discourse, in a unary manner, and this behaviour often conflicts with a plurality of truth properties.

I think one case where this issue is particularly apparent is in discourse which already involves the truth predicate. It is commonly assumed that 'truth' can be self-applicable; sentences involving the word 'true', or iterations of the word 'true', can be true or false. Some examples are: 'it is true that the sentence that $1+1=2$ is true', or: 'the sentence S is not true', where the sentence S is: '"grass is blue" is true'. It is of course well known that doing this naively (and often sophisticatedly!) can result in paradoxes, such as the Liar paradox,[6] but there are also many unproblematic sentences in which the truth predicate has self-applicable behaviour, such as the two sentences above.

Shapiro (2011) questions which domain of discourse these sentences should belong to, and how the pluralist can make sense of them. One option is for the pluralist to set aside a special alethic domain with its own alethic truth-property true_A, which all alethic sentences are true in virtue of satisfying. This appears to be an unsatisfactory answer, however, and falls into similar worries as above, where true_A appears to be adequate for all domains. We have the following equivalence: '"σ" is true_x' is true_A if and only if 'σ' is true_x. We also have the equivalence: 'σ' is true_x if and only if σ. Therefore we deduce 'σ' is true_A if and only if σ and no reference to true_x is required, no matter whether σ is a sentence from the domain of mathematics, morals, or minds. We are in the same issue as with a general truth property true_G, an alethic truth-property true_A would be sufficient as a truth property for any sentence σ.

It appears to me that the better option would be that if a sentence σ belongs to a domain with truth property true_x, then the truth property which ensures that 'σ is true' is true, is also true_x.[7] We have that 'σ' is true_x if and only if

[6] See Chapter 5 Section 5.2.5 where this paradox is introduced and discussed formally in relation to the first axiomatic theory of truth discussed in the chapter.

[7] This is somewhat similar to Lynch's (2013) response, where an alethic sentence is plainly true, and this supervenes upon the truth-manifesting property for the sentence which satisfies the truth predicate.

"'σ" is true' is true$_x$. This is a natural approach, but has issues with accounting for syntactically valid sentence like the truth-teller τ which is: 'τ' is true,[8] which appears then to have no truth property to make use of. This may not be an issue if domains of discourse are something like semantic natural kinds, where perhaps the truth-teller has no place, but for those like Lynch (2013) for whom domains of discourse are not rigid kinds, but simply a way to distinguish sentences based upon their logical form and content, this produces tension.

The pluralist needs to provide some kind of account for syntactically correct uses of the truth predicate, which does not conflict with her semantic account of truth properties. The problem for the pluralist is that the easy route to providing such accounts, adding in extra truth properties which can play this role, results in truth properties which can play *every* role. This is a contradiction in spirit, if not always in letter, with the pluralist motivation. The pluralist brings many theoretical virtues to the table, but the challenges I've considered above show a deep worry that the pluralist's position is not in accordance with natural ways that we use the truth predicate in ordinary language.

This worry is exacerbated by an objection known as the Quine-Sainsbury objection, which suggests that the alethic pluralist is not really talking about the property of truth at all. Sainsbury (1996), utilising an argument from Quine (1960) against pluralism about existence, argues that even if we recognise that '$1 + 1 = 2$' is true in a very different way to 'murder is wrong' is true, that this does not suggest that the truth property is different for each sentence. We can explain that these sentences are true differently, not because they make use of different alethic properties, but because numbers and wrongness are different things. Dodd (2013) contends that we do not need to bring in the meta-level concept of truth to explain these differences, when our object-level notions of the differences in content between the two sentences suffices. We can explain that '$1 + 1 = 2$' is true in a very different way to 'murder is wrong' is true, simply because the content of these sentences is very different. The Quine-Sainsbury objection questions why we should be focussing upon *truth* as the important factor, when it is content that can play this role just as easily. This is an implicitly deflationist reading of the pluralist about truth's motivation, but one with great weight.

In the next section I wish to develop this reading of the pluralist position into

[8]See Chapter 5 Section 5.2.5 again where the truth-teller is introduced and discussed in more detail.

an explicitly deflationary one. I argue that this theory provides an elegant combination of the two theories into a deflationary alethic pluralism, one which takes the explanatory benefits of a plural view of truth whilst avoiding these worries I've listed. I propose that if one finds the arguments for alethic pluralism convincing, then a deflationary alethic stance is a metaphysically innocent specification, which is well-motivated by our natural language use of the truth predicate.

6.3 Deflated Alethic Pluralism

For the purposes of this chapter I take a deflationary stance on truth to be one which recognises only a single insubstantial[9] property of truth, which is understood solely by an adequate account of the truth predicate's linguistic role. The truth predicate is understood primarily as a device of semantic ascent and descent, which enables one to take content from an object language, of subjects and predicates, and affirm it in a metalanguage - the subjects of which are the sentences of the object language, and vice-versa. One can take object-language level semantic content, e.g. that snow is white, and affirm it in a metalanguage by stating that the sentence 'snow is white' is true. This has two useful expressive purposes, as detailed by deflationists such as Horwich (1998). The first is that it enables one to affirm semantic content indirectly, e.g. the sentence S is true, where S could be 'the first sentence of this paragraph' or 'whatever I said this time last week'. The second is that it enables one to affirm a collection of content all in one go, e.g. all sentences satisfying P are true, where P might be 'being a mathematically proven theorem', or 'something Cassandra prophesied'. For some deflationists, this analysis of the truth predicate tells us all we need to know about truth, and all apparently deeper uses of truth can be explained by this account. The truth predicate is a useful linguistic device and expresses only a single insubstantial truth property which has neither metaphysical weight nor an explanatorily powerful role.

I wish to take this deflationary stance on truth as the starting point for my deflated pluralist position. Here I endorse only a very weak theory of truth, with the expectation that my remarks extend to stronger deflationary theories of truth, such as those introduced in Chapter 5. As sketched above, this position appears

[9]Quite what the claim of insubstantiality amounts to here is under debate. This is discussed in Chapter 3 where I argue that we should understand this to mean *pleonastic* in the sense of Schiffer (2003).

like a minimally adequate stance on the truth property, a single insubstantial property is ontologically more parsimonious than any number of substantive truth properties. Why might we say anything further about truth, then?

There are many reasons that pluralists, and advocates of more classical theories of truth, endorse additional metaphysical weight. For some like Wright (2001), we require a substantive theory of truth in order to have a satisfactory theory of assertoric content that holds across all domains of discourse. Others, such as Lynch (2013), view an account of truth as essential to explaining meaning and the norms of thought, which are invariant over the domains of discourse. Such challenges have been widely debated, and deflationists have provided counterarguments. For example Horwich (1998) has detailed many replies to many such objections. I do not wish to discuss their success or failure here, but instead respond to the sort of pluralist who views truth as essential to explaining key features of certain domains, and the deflationary picture as too minimal to do this. There are a number of philosophers (who are not all pluralists themselves) who make challenges such as these, which I will consider in more detail later. One example is that Shapiro (1998) argues that truth plays an (essential) explanatory role in mathematics, and deflationism cannot account for this. Boghossian (1990) argues that a deflationary conception of truth is unable to make sense of[10] non-factualism about a given domain of discourse, such as ethics. Asay (2009) argues that a constructive empiricist philosophy of science requires a substantive theory of truth, and hence deflationism cannot uphold such an account. In what follows I wish to challenge such views, and show that a deflationary alethic pluralism is capable of explaining these key features of certain domains, without requiring a substantive notion of truth.

I start with a language $\mathscr{L}$ and a truth predicate 'is true', where this predicate is understood purely as playing the role of semantic ascent and descent. The set of true sentences is then simply $\{\sigma : \sigma \text{ is an } \mathscr{L}\text{-sentence and } \sigma \text{ is true}\}$ and this is the extension of the truth property, the intension being the insubstantial property that the truth predicate expresses. This is the basis of the deflated alethic pluralist's theory of truth and all that they need say upon the matter.

The deflated alethic pluralist can, in addition to this, freely admit that there are other properties F_i which are extensionally equivalent to the truth property for

[10]In fact, Boghossian goes further and argues that deflationism is incompatible with non-factualism about a given discourse.

certain restrictions of the language $\mathscr{L}$. We might find there is a property F_i such that $\{\sigma : \sigma \text{ is an } \mathscr{L}_i\text{-sentence and } \sigma \text{ is } F_i\} = \{\sigma : \sigma \text{ is an } \mathscr{L}_i\text{-sentence and } \sigma \text{ is true}\}$ where $\mathscr{L}_i \subset \mathscr{L}$. In this instance, we can use the truth predicate interchangeably with F_i within the domain of discourse $\mathscr{L}_i$; F_i will also play a role of semantic ascent and descent, and might reasonably be called a 'truth-like' property (for $\mathscr{L}_i$).

Given a number of properties F_i which behave as above, we have a plurality of properties which are 'truth-like' for particular domains of discourse. It is important to note that these are not *truth* properties, however, but merely extensionally equivalent to truth properties for a particular range of sentences. There is no reason for these to be viewed as intensionally equivalent to truth properties, and in contrast to inflationary pluralism, there is only one genuine truth property. Truth is deflationary and monist, for there is only one property of truth, but this is not incompatible with admitting many properties which are equivalent to this truth property for certain domains of discourse.

Let us consider an example of this in (philosophical) practice. A sentence S is said to be superwarranted[11] if (and only if) it is warranted, and no matter how much further investigation takes place, will remain warranted. It has been suggested by Lynch (2009) that superwarrant is one of a plurality of truth properties, and that perhaps moral statements are true because superwarrant is the truth property for the moral domain. The deflationary alethic pluralist can also make a similar claim. Let us denote $\mathscr{L}_M$ as the domain of moral discourse and F_M as the property of superwarrant. The deflationist can happily claim that F_M is extensionally equivalent to the truth property for the domain of discourse $\mathscr{L}_M$. For moral sentences, the predicate 'superwarranted' can play a role of semantic ascent and descent, and expresses a truth-like property. Superwarrant is not, however, a truth property, since there are many sentences which could be true, but not superwarranted, perhaps 'the universe is infinite in area' is one such example. It might well be true, but certainly is not yet warranted, and so not superwarranted either.

An alternative example of this deflated alethic pluralism in action is in the domain of legal discourse. We might think that the truth property for a legal domain $\mathscr{L}_L$ is the property of coherence with the body of law. The deflated alethic pluralist can accommodate this claim by agreeing that $\{\sigma : \sigma \text{ is an } \mathscr{L}_L\text{-sentence}$

[11] This term originates from Wright (2001) who introduces the notion of superassertibility.

and σ coheres with the body of law$\} = \{\sigma : \sigma$ is an $\mathscr{L}_L$-sentence and σ is true$\}$, that cohering with the body of law is a 'truth-like' property for legal discourse. Whilst this does not make coherence with the body of law a truth property, it does allow the deflated alethic pluralist to formulate some useful claims about legal truth. For instance, perhaps the reason that legal truths are mutable over time is because the body of law is mutable over time. This is a feature of legal truth that the deflationary alethic pluralist can explain, which goes beyond her deflationary conception of truth.

Let us look at the challenges to deflationism I listed earlier. I aim to show that a deflationary alethic pluralism can overcome these. Shapiro (1998) argued that truth has an essential explanatory role in mathematical discourse, one which deflationism cannot access.[12] Shapiro remarks that, in the context of first order arithmetic, in order to explain why certain mathematical statements hold[13] we require the notion of truth (and more formally, we can prove these with an appropriate axiomatic theory of truth) and hence truth has an explanatory role in mathematical discourse. Shapiro suggests that a deflationist can avoid this argument by moving away from first order arithmetic and endorsing the view that arithmetical truth is a form of second-order or conceptual consequence. Shapiro then questions whether by accepting one of these notions, the deflationist would be adopting a robust notion of truth, since these appear to hide a rich concept of truth.The benefit of deflationary alethic pluralism is that it is able to endorse the view that second-order or conceptual consequence is a truth-like property for the mathematical domain, and that this can be used interchangeably with the truth predicate for mathematical sentences. These properties do not entail a robust notion of truth, since neither are identical to truth, but the deflationary alethic pluralist is able to acknowledge that within the mathematical domain the truth predicate is able to be used to phrase explanations within mathematics.

A similar claim can be made by the deflated alethic pluralist in response to Boghossian's challenge. Boghossian (1990) argued that deflationism cannot formulate non-factualism about a given domain (let's say ethics, for the sake of convenience). The reason for this is that a fact-stating sentence is one which is capable of being true or false, but Boghossian argues that for the deflationist this is purely a

[12]This is a philosophical explication of the conservativity argument which is discussed in detail in Chapter 4 Section 4.2.

[13]The Gödel sentence G is one such example – a sentence which is logically equivalent to 'G' is unprovable.

grammatical matter. Whether a sentence S is capable of being true or false, following semantic ascent and descent, becomes a question of whether S is well-formed or not. Boghossian concludes that, since ethical sentences are grammatical, they must trivially be fact-stating. The deflated alethic pluralist has room to respond here, and can claim that when we use the truth predicate to phrase non-factualism about ethics, we are actually using it as a substitute for the truth-like property E for the domain of ethics. This property E is extensionally equivalent with truth for ethical statements, but again, is not identical with it. The deflationary alethic pluralist can resist Boghossian's argument at its first step, and claim that ethical non-factualism is not really the claim that ethical statements are incapable of being true or false, but instead the claim that ethical statements are incapable of satisfying a certain truth-like property E, and that satisfying E is not a question of whether a sentence is well-formed or not. Whether such a response is convincing or not, it shows that the deflationary alethic pluralist has the resources to make claims within particular domains that substantivalists wish them to, without endorsing a substantive notion of truth.

The final challenge that I listed is from Asay (2009) who argues that deflationism can't make sense of a constructive empiricist approach to science. Constructive empiricism is the view that the aim of scientific inquiry is to produce empirically adequate theories, where empirical adequacy means that the observable features of observable objects the theory commits itself to are true. In order to make this claim, Asay (2009, p. 429) relies upon the understanding that "for the constructive empiricist, a theory is true just in case one of its models is isomorphic to the actual world" and that a deflationary theory of truth cannot make sense of this notion. Yet the deflationary alethic pluralist is an excellent position to make sense of this notion. They can make the claim that scientific truth is extensionally equivalent with having a model which is isomorphic to the actual world, and this is a truth-like property for the scientific domain. Deflationary alethic pluralism can happily make sense of constructive empiricist claims about science, and pairs extremely well with it. It enables the constructive empiricist to use the truth predicate as they wish, without smuggling metaphysical baggage into their theory of truth.

It is this explanatory feature of deflationary alethic pluralism that makes it an attractive position to hold. The deflationist can appropriate any explanatory power a substantive pluralist has about specific domains of discourse into her deflationary account of truth. The deflationary alethic pluralist can claim that any

truth property a pluralist has is a truth-like property, and explain any special features of truth for that domain utilising its truth-like property. For example, the deflationary alethic pluralist can claim that 'murder is wrong' will remain true in a hundred years time because 'murder is wrong' is superwarranted, and superwarranted sentences remain superwarranted over time. Perhaps this is different to the reason that mathematical truths will remain true in a hundred years time, and in opposition with legal truths, which may not be true in a hundred years time. The pluralist has access to a number of truth properties which give key explanatory benefits for certain domains, and the deflationary alethic pluralist has access to truth-like properties which perform a similar role. Further, the deflationary alethic pluralist can phrase such explanations using the truth predicate, instead of truth-like properties such as superwarrant, using the provision that this is restricted to a particular domain of discourse. This is no betrayal of the deflationist's basic stance that truth is not an explanatory notion, but allows the deflationist to easily explain particular features of certain truths.

It must be admitted that non-plural deflationists can perhaps explain these particular features of truths from specific domains as well. Deflationists often, by pointing to facts about certain statements within the object language, paraphrase away from 'truth'-talk in the metalanguage entirely. I have no qualms with this, but claim that the deflationary pluralist stance enables explanations which are commonly phrased with the truth predicate in the metalanguage, to remain in the metalanguage, without requiring a potentially problematic detour via the object language. This explanation avoids complicated paraphrase, and simply allows for the substantive pluralist's explanations using truth properties to be translated across to truth-like properties. The deflationary alethic pluralist position offers as much explanatory benefits to the theorist as a more substantive pluralist position, without adding weighty metaphysical properties in the mix. For the would-be pluralist, who is motivated away from deflationism by these considerations, deflationary alethic pluralism provides a more parsimonious theory which addresses these concerns.

A deflationary alethic pluralist still views truth in the simple and ontologically innocent way that other deflationists do. There is a unique truth property, which is understood by a deflationary conception of the truth predicate. We do not need weighty metaphysical notions to explain truth, and ascribe it no explanatory power beyond the predicate's linguistic role. We do, however, admit that there

can be truth-like properties which are explanatorily powerful, and which can play the explanatory role of particular truth properties for inflationary pluralists. We can think of the deflationary alethic pluralist as taking a 'best of both worlds' stance, where the parsimony of the deflationist is tied with the explanatory toolkit of the pluralist. I hope that this theory is philosophically adequate, in the sense of Chapter 4. I have shown that a deflationary theory of truth is able to accommodate philosophical uses of truth by endorsing additional truth-like (but not truth!) properties.

There is one area where one might object that additional weight has been smuggled in, however, and that is the truth-like properties and domains of discourse. I think there is a lot to be said here, and this shall be discussed in the next section.

6.3.1 Domains of Discourse and the Truth-Like Properties

In the previous section I have outlined a theory of deflated alethic pluralism, which I believe combines the best parts of both a deflationist and pluralist conception of truth. In this section I hope to provide some more details on this theory, and in particular the two under-specified aspects of this theory: domains of discourse and truth-like properties. It should be noted that the deflated alethic pluralist need not provide an account of either of these that is as comprehensive as the traditional pluralist requires. The deflated pluralist divorces usage of the truth predicate from the truth-like properties, and thus not every sentence need belong to a domain of discourse to be truth-apt. This means that the union of the domains of discourse need not be the entirety of the original language, and similarly the fusion of all the truth-like properties need not be extensionally equivalent with the truth property. Further, it is perfectly possible for a sentence to belong to multiple domains of discourse. As the deflated pluralist avoids conflating truth with the truth-like properties, a sentence could happily satisfy more than one of these properties. Even so, I would like to discuss a few stances that the deflated alethic pluralist can take on both domains of discourse and truth-like properties.

I hope that the theory, as currently sketched, is broad enough that almost any substantive pluralist attitude to domains of discourse can be adopted. It seems perfectly admissible to take domains of discourse as a prior theoretical notion, perhaps even as a primitive. These prior notions could be grounded simply in

intuition, since in most cases it appears that we do have an intuitive grasp of some domains of discourse such as: mathematical, moral, worldly, aesthetic, legal, etc. Given an account, intuitive or otherwise, of such pre-existing domains of discourse we could investigate if there are properties which are extensionally equivalent to the truth property within a certain domain, and take these as the truth-like properties. Whilst this appears a reasonable route for a deflationary alethic pluralist, it is not one that I personally would advocate. This brand of deflated pluralism is perhaps ensnared by the same issue that traditional plural views are by 'mixed' atomic sentences, as set out in Section 6.2.1. An account of domains of discourse resting upon intuition does not always result in an affirmative answer to which domain a sentence must belong to. The deflated alethic pluralist could consistently respond that mixed atomic sentences belong to no domain of discourse, but this seems unsatisfactory without a principled reason behind it. Further, any account of properties in terms of domains cannot guarantee that the truth-like properties such domains would generate would be the 'natural properties' that pluralists like to point to as candidates for truth properties. If the truth-like properties become things like 'approximately superwarrant aside from these cases, and with this addition' then the elegant explanatory power of the pluralist position is somewhat contaminated. That such truth-like properties do not occur would require further argumentation, and without a firm grasp on the domains of discourse, it seems hard to see how this could go.

Because of these reasons, my personal preference is to take the 'truth-like' properties as prior and generate domains of discourse from these. The deflated alethic pluralist can take already accepted philosophical notions, such as worldly truthmaking, coherence, superwarrant, etc. as truth-like properties and see to which class of sentences these properties are extensionally equivalent with the truth predicate. The domains of discourse would be these classes of sentences, closed under negation. This does not guarantee that these domains of discourse cohere exactly with our intuitive classification, but it seems reasonable to expect that there would be a loose matching, and this would be a useful method for precisifying these intuitive discursive areas. This has an added advantage over a 'domain first' pluralism that alternative philosophical notions can be investigated as potential truth-like properties without reprisal if they do not fit into already recognised domains of discourse. This 'property first' approach to deflated alethic pluralism does not rule out that some sentences will not be classified in any domain

of discourse, or even that they could be classified within multiple domains of discourse. This is not a problem for the view, since these sentences are true or false by the standard deflationary truth property, and I view such liberalness as a benefit.

If the deflated alethic pluralist takes this 'property first' approach, then there is an additional benefit that sentences which already involve the truth predicate are easily categorised. I raised this as an issue for the substantive pluralist in Section 6.2.1 and argued that a pluralist ought to take a sentence of the form 'σ is true' as belonging to the same domain of discourse as σ. The deflated alethic pluralist can happily accept that if σ is in a particular domain of discourse, then 'σ is true' also belongs in this domain. If σ satisfies a certain truth-like property, such as being superwarranted, then so will 'σ is true' as well. This is because the truth predicate is simply a logical-linguistic-semantic device that plays a role of semantic ascent and descent and doesn't affect the actual semantic status of the sentence, or any truth-like properties that it satisfies, in any way.

For the substantive pluralist, taking such a route had issue with accounting for which domain syntactically well-formed sentences like the truthteller τ, which is the sentence 'τ' is true, belong to, for such an account can offer no answer. The deflated alethic pluralist is in a good position to say that this sentence belongs to no domain of discourse, since the semantic properties of superwarrant, coherence, and other truth-like properties will not apply to purely syntactic sentences like τ. This is an example of how taking a deflationary pluralist stance can resolve one of the issues with standard inflationary pluralism I set out in Section 6.2.1. In the next section I aim to show that a deflationary stance on truth can help to resolve the other issues that I set out as well, and thus deflationary pluralism is motivated by not just parsimony grounds, but also in its ability to resolve standard challenges to alethic pluralism.

6.3.2 Resolving Pluralism's Issues

I have argued that deflated alethic pluralism has a good approach to sentences involving the truth predicate, but what about the other objections to pluralism that I listed during Section 6.2.1? Many of pluralism's issues are with accounting for sentences which are not easily categorised within a single domain of discourse, for instance mixed atomic sentences, mixed compound sentences, and general sen-

tences. Happily, the deflated alethic pluralist is able to block all of these issues immediately. For the deflated alethic pluralist, a compound, mixed atomic, or general setence 'S' is true if and only if S. There is nothing further the deflated alethic pluralist needs to say here, since the truth-like properties do not tell us anything about the insubstantial truth property, beyond a fragment of its extension and anti-extension. The deflated alethic pluralist is not concerned by their ability to talk about the truth of such sentences, since deflationism is adequate for explaining their behaviour. Further, the deflated alethic pluralist need not concern themselves with the domain of discourse such sentences reside in. These sentences could belong to no domain of discourse, or many, and neither is problematic for the deflationary alethic pluralist, since these are untethered from questions of truth.

One of the more pressing challenges to pluralism about truth is the Quine-Sainsbury problem, which questions pluralism's foundational motivation. The objection contends that pluralism's defining feature, that different sentences are true in different ways, questions why *truth* should be different for these sentences, when the object-level notions employed by such sentences suffices. Deflated alethic pluralism embraces this reading of pluralism. *Truth* is not different for such sentences, there is one truth property, and it is used in the same way for every such sentence. The object-level notions sentences use are satisfactory to explain these differences, and often in quite a general way. It is these object-level notions that give rise to the truth-like properties, which do the explanatory work here. My deflated pluralist stance deflates the pluralist position to one where a plurality of truth properties are not needed, for the truth-like properties can perform this job just as adequately.

I conjecture that the deflationist stance on the linguistic purpose of the truth predicate, as a generalisation device, explains how the pluralist reaches this inflated position of plural truth properties. We often use the truth predicate to discuss a general group of sentences in one go, for example 'if a sentence expresses the conclusion of a sound argument, then it is true'. The deflationist is quick to point out that we are not, really, remarking on a feature of *truth* in this example, but a feature of conclusions of sound arguments. We do express this feature using this generalising truth predicate, however. In many linguistic contexts there are certain special features of true sentences. For instance, perhaps a mathematical sentence is true if it can be proven. We use the truth predicate to express this general feature of mathematical sentences, giving an illusion of a special truth

property for mathematical sentences which makes this so. The deflationary line is that this sentence is not about mathematical truth, though, but mathematical proof. Truth is not the subject of interest here, and is merely a helpful way to phrase this interesting feature of proven mathematical sentences. Mathematical proof is certainly an interesting concept, and the pluralist is correct to observe that properties which play a similar role across other discursive contexts are of philosophical interest too. It seems incorrect to claim these are truth properties, though, and perhaps this error results from mistaking the truth predicate's role of generalisation as indicative of something deeper. In the example above it is the notion of proof that is deep and interesting, and the concept worthy of inquiry, not truth.

A deflated pluralist stance moves pluralism away from difficulties concerning usage of the truth predicate, and issues of truth altogether, and into the more fruitful starting point of investigating these properties which, in certain domains, are extensionally equivalent with the insubstantive truth property. These properties certainly appear to have interesting metaphysical, epistemic, and normative features, which are worthy of investigation, and are at least (and at most!) extensionally closely tied to truth. The pluralist framework misconstrues these as truth properties which raises a host of problems and objections. By taking a deflated alethic pluralist stance, these problems are easily avoided, and fruitful research on these properties is untethered from remarks on truth.

6.3.3 Distinguishing Deflated Alethic Pluralism

I have argued that a deflated pluralist theory provides a framework for researching truth-like properties which avoids issues besetting substantive pluralist theories of truth. It might be thought that this proposal is, perhaps, merely terminologically distinct from the other alethic pluralisms, however. Perhaps I have simply made a formal manoeuvre which, predominantly, just renames truth properties as truth-like properties and differs to current pluralisms only in its presentation. In this section I wish to push back against this impression, and argue that deflated alethic pluralism is a genuinely distinctive position with sizeable benefits over-and-above sidestepping certain issues besetting pluralisms about truth.

I hope it is clear that my position is ontologically lighter than Wright's (1998) pluralist stance, where there are many properties of truth. I admit only one

property of truth, and this is insubstantial. Edwards' (2011) Simple Determination Pluralism is more similar to my view, as it admits only a single property of truth, but requires that the truth-determining properties in domains determine whether a given sentence is true or not. In opposition to Wright and Edwards, I do not require that a sentence has to satisfy any truth-like property in order to be true, or to even belong to a (unique) domain of discourse. Another key distinction is that Wright and Edwards understand truth via a list of platitudes, features all truth properties satisfy, but such lists are lengthy and contentious. On the other hand, I understand truth simply via a deflationary account of semantic ascent and descent, and no investigation or acknowledgement of further platitudes is required.

Deflated alethic pluralism is simpler than Lynch's (2009) functionalist theory of truth in these regards as well. Lynch's truth functionalism may not require a plurality of truth properties to fulfil the truth role, but he does treat truth as a functional kind that can be realised in many different ways. This functional kind is described by certain truisms, and Lynch leaves it open whether his list is complete or some of these truisms might be replaced. A key distinction between deflationary alethic pluralism and Lynch's account is that I offer a simple account of truth without requiring investigation of which truisms are the correct role-descriptions of truth. Further, treating truth as a functional kind can lead to significant issues. Wright (2013) argues that Lynch's manifestation-functionalism characterisation of truth is self-refuting, since one of the essential truisms will be that truth is manifested by the various properties which manifest it, but these properties cannot manifest each other.[14] I avoid such formulation issues entirely, since *truth* bears no relationship (other than part extensional equivalence) with the truth-like properties. Further, treating truth as a functional kind might well still be ontologically more burdensome than a deflationary conception of truth. Hiddleston (2011) argues[15] that manifestation-functionalism is committed to second-order properties - a commitment that deflationary alethic pluralism avoids. A truth pluralism relying upon functionalism is at a disadvantage when compared to my deflationary alethic pluralism.

Such concerns motivate my account over more deflationary variants of functionalism as well. One titularly similar direction is Edwards' (2012) Deflationary

[14]Lynch (2013) responds to this challenge, but Wright (2013) remains unconvinced.

[15]Hiddleston here is talking about manifestation-functionalism within the context of philosophy of mind, but I see no reason that his argument and examples cannot be adapted to the alethic case.

Functionalist theory of truth,[16] which modifies Lynch's account to a deflationary pluralism. In Edwards' deflationary functionalism, each sentence plays the truth-role for itself, and there is a truth property for every sentence. This is a deflationary functionalist position, since we have no common truth properties, but a plurality of deflated properties. In deflationary functionalism the truth-role is still provided by core truisms, and thus this theory is not wholly deflationary (since it treats the norm of belief as a core truism) but we could go further and consider a minimal functionalist theory of truth. Borrowing Horwich's (1998) minimal theory of truth, we could take Edwards' deflationary functionalism with the only truism being that 'P' is true if and only if P. This certainly appear to be a genuine deflationary version of pluralism, but still falls into issues with treating truth as a functional kind. Further, deflationary pluralisms such as this still require a property to play the truth-role for every sentence. It is a key benefit of my deflationary alethic pluralism that not every sentence needs a unique truth-like property in order to be true, and this has substantive benefits when avoiding issues of classifying mixed atomic sentences into a unique domain of discourse.

One last position which terminologically might be thought similar to my theory is Beall's (2013) Deflated Truth Pluralism. Beall introduces a view where the deflationist can admit a plurality of truth properties. Beall advocates that a deflationst can accept different deflationary truth properties for different languages which utilise different logics. This view is not incompatible with my view, but significantly different. I start with one deflationary truth property for one language with one logic and do not consider alternative languages and logics. Our views are not incompatible, far from it, but the focus is substantively different: I look at subsets of one language, whereas Beall explores multiple languages.[17] Beall has no 'truth-like' properties which play a role beyond semantic ascent and descent, and particularly none which do this for a subset of the language.

[16] I do not wish to imply that Edwards advocates this theory of truth, it is merely introduced as a possible theory to challenge Lynch's claim that his theory is significantly more substantive than a deflationary account.

[17] Beall does note a possible view 'language relative truth pluralism' which introduces truth predicates for fragments of a language, which appears much more similar to my view. This position is only noted, and Beall discusses 'language wide truth pluralism' instead.

6.4 Conclusion

I hope that these reflections show that my view is genuinely distinctive from current pluralist positions, but one with significant attraction to the would-be pluralist. By untethering the truth predicate from properties which can play a truth-like role, a deflated alethic pluralist enjoys the benefits of both a deflationary and pluralist position. The difficulties facing pluralism about truth can be avoided by establishing a distinct theory of the truth predicate, which behaves in a monist deflationary way. This results in a monist deflationary truth property, which is strictly speaking all the deflationary alethic pluralist says about truth. The deflationary alethic pluralist can admit multiple properties which are, for certain subsets of the language, extensionally equivalent with the truth property. These properties can be as substantive and explanatory as needed, providing a deflationary alethic pluralist with all the rich theoretical structure of interest to the pluralist.

I have established the framework of a deflated alethic pluralist theory, one which perhaps allows the problems besetting pluralism to be moved on from, and hopefully enabling further research in what the interesting truth-like properties are and how they behave. I postulate that interesting contenders for truth-like properties are the current contenders for substantive pluralist truth properties; properties such as coherence, superwarrant, correspondence, etc. It can be investigated for which sentences these properties are extensionally equivalent with a deflationary truth property, and whether the resulting class of sentences, when closed under negation, approximates an intuitive domain of discourse. An alternative approach would be take to an intuitively plausible domain of discourse and attempt to 'reverse engineer' a natural truth-like property which approximately generates it. It can then be seen how these truth-like properties behave, and what interesting explanatory power and metaphysical contributions they can bring to the table. This would specify the deflated alethic pluralist theory beyond the general formulation I have given it here, and highlight the benefits this account has over a traditional non-augmented deflationary theory of truth.

These are questions left open for further research, however. My aim in this chapter has been to provide a deflationary theory of truth which can admit the same explanatory strength as a pluralist theory of truth. Plural theories of truth are able to provide for all uses of 'truth' in different domains of discourse, because

of their numerous substantive truth properties, and thus should be seen as philosophically adequate. I therefore take it that deflationary theories of truth can be philosophically adequate as well. I have shown that even a very weak deflationary theory of truth, admitting only that the role of the truth predicate is semantic ascent and descent, is able to do this. I hope that, hence, a strong deflationary theory such as Axiomatic Typed Truth (ATT), introduced in Chapter 5, is able to do this as well. This answers the concluding question of Chapter 4 and proposes that an axiomatic (deflationary) theory can be both formally and philosophically adequate. This answers one of the motivating question of this thesis, to be argued for properly in the conclusion, Chapter 7, next, that axiomatic theories of truth support deflationism about truth. This leads me to conclude that a deflationary conception of truth is correct, again to be argued for in the following chapter.

Chapter 7

Conclusion

I began this thesis by asking what our concept of truth is. I have been particularly interested in what the behaviour, nature and role of the truth property is. I have focused this question by exploring what a deflationary answer to these questions is and whether this answer is adequate. My methods have been partly formal and one key question has been whether formal theories of truth support deflationism about truth. My answer is that a deflationary view of truth is adequate. I claim that the research contained in my thesis supports the view that an axiomatic view of truth is adequate for formal and philosophical purposes and hence supports deflationism about truth. In the following section (Section 7.1) I shall provide a reminder of some of the main contributions of this thesis, and their importance, and then in Section 7.2 I shall demonstrate how this conclusion follows and its wider significance. I end with Section 7.3 in which I look at further questions and research that have been inspired by this work.

7.1 Summary of Thesis

In Chapter 2 I developed an extended T-schema for nonstandard models of syntax. I proposed this as a new minimal adequacy condition for theories of truth and showed that closing CT^- (Compositional Truth without induction axioms) under this schema results in a non-conservative theory of truth over arithmetic. This provided a novel strengthening of the conservativity argument against deflationism, responding to Field's (1999) counterargument, by showing non-conservativity of truth from purely alethic considerations. This leads to a choice for the deflationist:

they can accept conservativity, and argue against nonstandard models of syntax, or they can deny conservativity and boast the deductive power of an (extended) T-schema and compositional clauses. In answer to this, I argued in Chapter 4 that the deflationist should deny conservativity.

Chapter 3 questioned what a deflationary theory of truth is and presented a new understanding of deflationism as a logical-linguistic-semantic theory of 'true'. This clarifies 'deflationism' (as it applies to truth) as a term of art and I provided a criterion of deflationism which accords with our current usage. The theories it categorises as deflationary are those which are labelled as such. Further, I argued that deflationary truth properties are *pleonastic*, in the sense of Schiffer (2003), and showed this is an improvement to alternative understandings of what it means for truth properties to be 'insubstantial'. My conception of deflationary theories of truth rescues deflationism from being equated with a T-Schema, an equation I argued against in the chapter. This means that many existing challenges to deflationism are only really challenges to the T-Schema as a theory of truth and we should not view deflationism as threatened by these.

Chapter 4 critically discussed which formal theories of truth are deflationary and concluded that all axiomatic theories of truth are deflationary. This followed from my philosophical criterion of deflationism in Chapter 3. I provided novel criticisms of conservativity arguments against deflationism and a new examination of the 'logicality' of truth. This has provided a new defence to deflationists from these formal challenges and showed that work in formal theories of truth can support deflationism. I concluded that a primary question of interest is whether an axiomatic theory of truth can be adequate formally and philosophically, since if so we would have an adequate deflationary theory of truth.

In Chapter 5 I developed and explored two new axiomatic theories of truth. I introduced a new axiomatic typed theory of truth (ATT) as an improvement over existing typed theories of truth and discussed its interesting application to questions over semantic paradoxes and absolute generality. This inspired a new type-free theory of truth with interesting connections with the Kripke-Feferman (KF) theory of truth. I remarked on the connection between these theories and concluded that ATT should be seen as formally adequate. Both of these new theories of truth still have many interesting open questions surrounding them and inspire research developing and investigating similar theories of truth, in particular in the type-free case.

Chapter 6 advanced a deflationary theory of truth that can incorporate the explanatory benefits of a pluralist theory of truth. I showed how the deflationist can endorse a metaphysically insubstantial, but explanatorily powerful, theory of truth by appealing to truth-like properties – properties extensionally equivalent to truth for particular domains of discourse. I argued that this shows even a weak deflationary theory of truth can be philosophically adequate and that this theory overcomes key contemporary challenges to standard deflationary theories of truth. This shows that a deflationary view of truth can be adequate for philosophical purposes and inspires questions over what exactly these truth-like properties and their corresponding domains of discourse are.

Whilst these are the main contents of the chapters of this thesis, and of interest individually, together they provide a cohesive argument for deflationism about truth. This will be the subject of the next section, where I present this argument in detail.

7.2 A Defence of Deflationism

In this section I will bring my thesis together and provide the reason that I believe a deflationary conception of truth is correct, based on the research within this thesis. I view a deflationary conception as the *default* conception of truth. A deflationary position is ontologically lighter than all competing theories of truth,[1] and thus as the default position, it only needs to be shown that the theory is adequate explanatorily. I shall provide a defence that deflationism has enough explanatory power, and therefore an argument for deflationism about truth.

This exploration of the explanatory power of a deflationary theory of truth needs unpacking. One of the central tenets of a deflationary conception of truth is that the *property* of truth lacks in 'causal-explanatory' power. This is an important difference to the claim that a *theory* of truth should lack in explanatory power. A theory of truth ought to be able to explain the nature and behaviour of truth and detail how this is exemplified by the role of the truth predicate. A deflationary

[1]This is not quite true, as it might be argued that a redundancy theory of truth is ontologically lighter still. This is a theory, often ascribed to Ramsey (1927), which proposes that the truth predicate is redundant linguistically and that 'p is true' means the same as 'p'. This is perhaps even more minimal than deflationism, since it posits *no* truth property, but does not appear adequate in light of linguistic generalisations, detailed in Chapter 3 Section 3.2, which appear to be a counterexample to this theory.

theory of truth is no different, and ought to be able to explain these features as well. What makes deflationism special is that it claims that the truth property has no substantial nature and no 'causal-explanatory' power. This is hard to reconcile with the role of the truth predicate, which is frequently used to phrase explanations in philosophy and beyond. The challenge for deflationism is to provide an adequate description of the role of the truth predicate, accurate to its suitable usage, in which explanations using the truth predicate do not require a substantive or explanatory truth property. In this thesis I have aimed to provide arguments to this effect.

Chapter 3 provided a conceptual analysis of what it means to be a deflationary theory of truth and clarified what the term 'insubstantial' means. I concluded that a deflationary theory of truth is a logical-linguistic-semantic theory of word 'true' and that a deflationary property of truth is *pleonastic* in the sense of Schiffer (2003). This led to the conclusion of Chapter 4 that all axiomatic theories of truth are deflationary theories of truth. It thus follows that if it can be shown we have an axiomatic theory of truth which is adequate to explain the role of the truth predicate in phrasing explanations, which coheres with suitable usage of the predicate, then we have an adequate deflationary theory of truth. If a deflationary theory of truth is adequate, then as the *default* conception of truth, deflationism should be held as the correct conception of truth.

By suitable usage I mean something a little more than 'formal adequacy' in the sense of Chapter 4 Section 4.5, and a little less than 'philosophical adequacy', also discussed there. A theory should respect the way we use the truth predicate in natural language modulo certain normative considerations – the primary consideration being consistency. This means that a theory should have as consequences as much of a T-Schema and compositional clauses as possible. This T-Schema should not be restricted to standard models of syntax and entail an extended T-schema, in the sense of Chapter 2, for the reasons argued therein. The theory should also endorse only a single truth predicate and this predicate should be 'self-applicable', so that we can predicate truth of sentences already containing the truth predicate. Further, the theory should be phrased within a classical metatheory. Given such desiderata, the theory ATT introduced in Chapter 5 Section 5.2, appears to cohere with suitable usage of the truth predicate.

The theory of Axiomatic Typed Truth classifies sentences as 'truth-apt' or 'not-truth-apt'. For those sentences which are 'truth-apt', the theory entails a full T-Schema, an extended T-schema when considered over nonstandard models

of syntax, and full compositionality. The theory endorses a single truth predicate which can be used self-applicably and is formed within a usual classical metatheory. Most importantly, the theory is consistent, due to its classification of the paradoxes as 'not-truth-apt'. The theory is typed in nature, but this can be defended by appeals to a contextual notion of truth, that the word 'true' is relativised to contexts of use. Further, sentences which quantify absolutely over all levels of truth are classified as 'not-truth-apt', which also fits with a contextual notion of quantification. The theory's truth predicate is close to natural language usage of the truth predicate, whilst still retaining consistency, and the trade-offs made for this have suitable philosophical defence. I therefore claim that ATT is an adequate axiomatic (and thus deflationary) theory of truth in the sense that it coheres with suitable usage of the truth predicate.

Do we therefore have an adequate deflationary theory of truth? I believe so. This follows from my arguments in Chapter 6 which advanced that a weak deflationary theory of truth is adequate at providing for the role of the truth predicate in phrasing explanations. This is philosophical adequacy in the sense of Chapter 4. The truth predicate is used in many areas of philosophy to explain key features of certain sentences. In the introduction to Chapter 4 I provided a sample of these: knowledge is justified true belief, logical connectives' meanings are given by their truth conditions and the aim of science is truth. The theories of truth most able to explain the role of the truth predicate in these areas of philosophy and more are pluralist theories of truth. These theories endorse multiple 'domains of discourse' each of which has an individual truth property, fitted to provide the most appropriate explanatory uses of truth for that domain. I argued in Chapter 6 that even a weak deflationary theory of truth (a consistent T-Schema) is able to appropriate this explanatory power, however. This means that an axiomatic theory of truth which entails a T-Schema, such as ATT, can explain the role of the truth predicate in phrasing these explanations. We therefore have an adequate deflationary theory of truth.

I have aimed to defend deflationism from those who would claim that deflationary theories of truth do not have enough explanatory power. The theory ATT is deflationary, since it is axiomatic and thus logical-linguistic-semantic in nature. Further, this theory provides for key philosophical uses of the truth predicate and is accurate to the predicate's suitable usage. Therefore, I claim that we have an adequate deflationary theory of truth and hence deflationism should be held as the

correct conception of truth.

The importance of this claim is as far-reaching as it is modest. If deflationism about truth is correct, then those seeking *truths* do not need to be interested in *truth*. When the scientist conducts experiments to ascertain truths of the world or the mathematician produces proofs to discover mathematical truths, they are not doing this to seek *truth* itself. I am simply using the truth predicate to phrase their endeadvour, perhaps as a device of generalisation to express truth-like properties for their respective domains, as suggested in Chapter 6. Similarly the court of law determining a charges' truth or the stockbroker ascertaining the truth of a forecast are not seeking *the truth*, but particular *truths*. I use the truth predicate as a device of 'blind ascription' to phrase these individual inquirations – the predicate is used to affirm a sentence which has been named rather than quoted. In each of the examples given at the start of this thesis, *truth* is not the focus, but *truths*. The search for these is phrased with a truth predicate, but the pleonastic property of truth, provided by a linguistic transformation of statements involving the truth predicate, is not of primary importance.

This is not to say that the *truths* searched for are not important, but that the question as to whether and why they are important is not a question about truth. Such questions can be phrased using the truth predicate, but this is to use it as a device of generalisation. In fact, we may think that this is using the truth predicate as a device of *over*-generalisation, for without undue precision of domains it can hide distinctions between the truth-like properties. Whether a true sentence is important depends upon the particular semantic content that the sentence expresses, rather than the nature of truth itself. This means that those who seek truths are seeking nothing about the property of truth other than a fragment of its extension.

I view this thesis as providing consolation to those who seek truths. The experts of truth for a particular domain are not experts of theories of truth, but experts of that domain. Seekers of truths do not need to concern themselves with general metaphysical or epistemic notions inherent to the nature of truth, particularly any which could offer conceptual or practical conflict with their current practice. Nor, for sentences within their respective domain of discourse, do they even need to concern themselves with a complex theory of truth such as ATT, for most sentences are true or false by a simple typed T-Schema. Really, they do not need to concern themselves with *truth* at all. They can continue seeking their particular

semantic content and continue to phrase such investigations and the result of such investigations with the truth predicate, however.

Whilst this leaves important questions of *truth* itself answered, this leaves many more questions open and raises a host of new ones. I shall discuss such questions in the following section and suggest the direction of further research that is inspired by this thesis.

7.3 Further Research

My research has inspired a variety of further questions, both formal and philosophical, on the subject of truth and philosophy more generally. I have provided a number of such questions and suggestions throughout the thesis and proposed specific open mathematical problems which have been summarised at the end of Chapter 2 and Chapter 5. My aim in this section is not to repeat these questions, but to highlight the wider theme which recurs throughout the thesis and future directions that research could take in light of this.

The research project in this thesis has been an exploration of deflationism about truth and whether formal theories of truth support this or not. The natural extension of this research is to look at deflationism in philosophy more generally, whether we have formal theories of such concepts and, if so, whether they support deflating that particular topic. There is much contemporary interest in deflationary theories of metaphysics, content and reference beyond deflationism about truth. Thomasson (2015) champions a deflationary approach to ontology, Field (1994a) argues for a deflationary theory of content and Båve (2009) proposes a deflationary theory of reference, for example. It is natural to consider whether investigations analogous to the research contained in my thesis could be carried out for these theories as well.

The first question that could be addressed is whether these topics are deflationary in the same sense as theories of truth are deflationary. I argued in Chapter 3 that the term 'deflationary', as it applies to truth, describes a logical-linguistic-semantic theory of the word 'true'. Is this usage consistent across philosophy, or does the term 'deflationary' vary in its meaning? For example, is a deflationary theory of ontology a logical-linguistic-semantic theory of the word 'exists', or does this not characterise deflationism about ontology? Such theories usually 'deflate' the term to a quasi-logical notion and it would be interesting to see whether this

is similar to the truth-deflationists' use of 'quasi-logical'.

Given that these deflationary approaches are often described as quasi-logical, it seems possible, if these theories are consistent, to develop formal theories of them. Given deflationism about X, can we develop semantic and axiomatic theories of X? Work in this area seems in its infancy and ripe for development. As a starting point, Zalta (1983) has developed an axiomatic theory of metaphysics and Picollo (2018) has recently proposed a formal theory of reference. Are these deflationary theories of metaphysics and reference, respectively, and can further theories be developed?

Formal theories of truth could be a useful aid in developing these formal theories of other topics. In Chapter 5 I suggest that formal theories of truth could be a useful research tool for questioning the correctness of absolutely general quantification. Formal truth theory could be used to develop new theories of this, and also other areas of philosophy. As an example of this, Stern (2014a,b) has used axiomatic theories of truth to develop formal theories of modality. Perhaps Stern's theories can be used to investigate deflationism about modality?[2] Given such formal theories, one natural extension of my research is to consider the adequacy of these over nonstandard models, similar to my research in Chapter 2. It appears that my arguments in Chapter 2 for the relevance of nonstandard models to studies of truth would generalise to other areas of philosophy as well.

If we could construct a formal theory of such notions, conservativity arguments might be brought to bear upon them. Schiffer (2003, p. 56), for instance, argues that pleonastic concepts[3] are something like proof-theoretically conservative concepts. It would be interesting to see whether such formal theories could be conservative and the philosophical importance of this. I conjecture that my arguments against the truth-conservativity argument in Chapter 4 would extend beyond truth and count against substantive philosophical importance of conservativity, however. Perhaps a better test, as suggested in Chapter 4, is whether we can provide a formal test of quasi-logicality for such theories. We have a number of proof-theoretic and semantic tests of 'logicality', but it is open whether these can be extended to broader notions, such as being logical-linguistic-semantic. If so, it seems that we would have a clear test of whether a particular formal theory

[2]Sidelle (1989) appears to advocate something like a deflationary conception of modality, for example.

[3]Similar to a pleonastic property, a pleonastic concept in general is one which results from a 'something-from-nothing' linguistic transformation.

is deflationary or not. Such a test could also settle whether any semantic theory of truth is deflationary, a question currently left open by my research.

These questions offer a wealth of further research to consider, but the primary question of this thesis has been given an answer. The nature, role and behaviour of truth is deflationary and this is supported by research in formal theories of truth. Therefore, a deflationary concept of truth is correct.

Bibliography

Noga Alon and Joel Spencer. *The Probabilistic Method*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2016.

Bradley Armour-Garb. Deflationism (about theories of truth). *Philosophy Compass*, 7(4):267–277, 2012.

Bradley Armour-Garb and Jc Beall. *Deflationary Truth*. Open Court Press, 2005.

Jamin Asay. Constructive empiricism and deflationary truth. *Philosophy of Science*, 76(4):423–443, 2009.

Jamin Asay. Primitive truth. *Dialectica*, 67(4):503–519, 2013.

Jody Azzouni. Truth via anaphorically unrestricted quantifiers. *Journal of Philosophical Logic*, 30(4):329–354, 2001.

Jody Azzouni. *Tracking Reason: Proof, Consequence, and Truth*. Oxford University Press, 2006.

Arvid Båve. A deflationary theory of reference. *Synthese*, 169(1):51–73, 2009.

Jc Beall. *Spandrels of Truth*. Oxford University Press, 2009.

Jc Beall. Deflated truth pluralism. In Nikolaj Pedersen and Cory Wright, editors, *Truth and Pluralism: Current Debates*, pages 323–338. Oxford University Press, 2013.

Jc Beall and Bradley Armour-Garb. *Deflationism and Paradox*. Oxford University Press, 2005.

Paul Boghossian. The status of content. *Philosophical Review*, 99(2):157–84, 1990.

Leonard Bolc and Piotr Borowik. *Many-Valued Logics 1: Theoretical Foundations*. Springer Verlag, 1992.

Denis Bonnay. Logicality and invariance. *Bulletin of Symbolic Logic*, 14(1):29–68, 2006.

Denis Bonnay and Henri Galinon. Deflationary truth is a logical notion. In Mario Piazza and Gabriele Pulcini, editors, *Philosophy of mathematics: Truth, Existence and Explanation*, pages 71–88. Springer, 2018.

George Boolos, John Burgess, and Richard Jeffrey. *Computability and Logic*. Cambridge University Press, 2007.

Robert Brandom. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, 1994.

Robert Brandom. Expressive vs. explanatory deflationism about truth. In Richard Schantz, editor, *What is Truth?*, pages 103–119. Walter de Gruyter, 2002.

Tyler Burge. Semantical paradox. *Journal of Philosophy*, 76(4):169–198, 1979.

Andrea Cantini. Notes on formal theories of truth. *Mathematical Logic Quarterly*, 35(2):97–130, 1989.

C. T. Chong, Wei Li, and Yue Yang. Nonstandard models in recursion theory and reverse mathematics. *The Bulletin of Symbolic Logic*, 20(2):170–200, 2014.

Cezary Cieśliński. Deflationism, conservativeness and maximality. *Journal of Philosophical Logic*, 36(6):695–705, 2007.

Cezary Cieśliński. Truth, conservativeness, and provability. *Mind*, 119(474):409–422, 2010a.

Cezary Cieśliński. Deflationary truth and pathologies. *Journal of Philosophical Logic*, 39(3):325–337, 2010b.

Cezary Cieśliński. The innocence of truth. *Dialectica*, 69(1):61–85, 2015.

Cezary Cieśliński. *The Epistemic Lightness of Truth: Deflationism and its Logic*. Cambridge University Press, 2017.

Tim Crane. *The Objects of Thought*. Oxford University Press, 2013.

Nic Damnjanovic. New wave deflationism. In Cory Wright and Nikolaj Pedersen, editors, *New Waves in Truth*, pages 45–58. Palgrave Macmillan, 2010.

Marian David. The correspondence theory of truth. In Michael Glanzberg, editor, *The Oxford Handbook of Truth*, pages 238–258. Oxford University Press, 2018.

Walter Dean. Models and Computability. *Philosophia Mathematica*, 22(2):143–166, 2013.

Julian Dodd. Deflationism trumps pluralism! In Nikolaj Pedersen and Cory Wright, editors, *Truth and Pluralism: Current Debates*, pages 298–322. Oxford University Press, 2013.

Douglas Edwards. Simplifying alethic pluralism. *Southern Journal of Philosophy*, 49(1):28–48, 2011.

Douglas Edwards. Alethic vs deflationary functionalism. *International Journal of Philosophical Studies*, 20(1):115–124, 2012.

Douglas Edwards. Truth as a substantive property. *Australasian Journal of Philosophy*, 91(2):279–294, 2013a.

Douglas Edwards. Truth, winning, and simple determination pluralism. In Nikolaj Pedersen and Cory Wright, editors, *Truth and Pluralism: Current Debates*, pages 113–122. Oxford University Press, 2013b.

Matti Eklund. What is deflationism about truth? *Synthese*, 2017. Published Online First: https://doi.org/10.1007/s11229-017-1557-y.

Ali Enayat and Fedor Pakhomov. Truth, Disjunction, and Induction. *ArXiv e-prints*, 2018. arXiv:1805.09890.

Fredrik Engström. Satisfaction classes in nonstandard models of first-order arithmetic. Master's thesis, Chalmers University of Technology and Göteborg University, 2002.

Solomon Feferman. Systems of predicative analysis. *Journal of Symbolic Logic*, 29 (1):1–30, 1964.

Solomon Feferman. Reflecting on incompleteness. *Journal of Symbolic Logic*, 56 (1):1–49, 1991.

Solomon Feferman. Set-theoretical invariance criteria for logicality. *Notre Dame Journal of Formal Logic*, 51(1):3–20, 2010.

Hartry Field. Critical notice: Paul Horwich's 'Truth'. *Philosophy of Science*, 59 (2):321–330, 1992.

Hartry Field. Deflationist views of meaning and content. *Mind*, 103(411):249–285, 1994a.

Hartry Field. Disquotational truth and factually defective discourse. *Philosophical Review*, 103(3):405–452, 1994b.

Hartry Field. Deflationist views of meaning and content. *Mind*, 103(411):249–285, 1994c.

Hartry Field. Deflating the conservativeness argument. *Journal of Philosophy*, 96 (10):533–540, 1999.

Hartry Field. *Saving Truth From Paradox*. Oxford University Press, 2008.

Martin Fischer and Leon Horsten. The expressive power of truth. *Review of Symbolic Logic*, 8(2):345–369, 2015.

Harvey Friedman and Michael Sheard. An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33(1):1–21, 1987.

Kentaro Fujimoto. Deflationism beyond arithmetic. *Synthese*, 196(3):1045–1069, 2019.

Henri Galinon. *Recherches sur la vérité – Définition, élimination, déflation (Investigations on the notion of truth – definition, elimination, deflation)*. PhD thesis, Université Paris-1 Pantheon Sorbonne, 2010.

Henri Galinon. Deflationary truth: Conservativity or logicality? *Philosophical Quarterly*, 65(259):268–274, 2015.

Michael Glanzberg. The liar in context. *Philosophical Studies*, 103(3):217–251, 2001.

Michael Glanzberg. Quantification and realism. *Philosophy and Phenomenological Research*, 69(3):541–572, 2004.

Nelson Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, 1955.

Patrick Greenough. Deflationism and truth-value gaps. In Cory Wright and Nikolaj Pedersen, editors, *New Waves in Truth*, pages 115–125. Palgrave-Macmillan, 2010.

Dorothy Grover, Joseph Camp, and Nuel Belnap. A prosentential theory of truth. *Philosophical Studies*, 27(2):73–125, 1975.

Anil Gupta and Nuel Belnap. *The Revision Theory of Truth*. MIT Press, 1993.

Volker Halbach. Tarskian and kripkean truth. *Journal of Philosophical Logic*, 26 (1):69–80, 1997.

Volker Halbach. Disquotationalism and infinite conjunctions. *Mind*, 108(429): 1–22, 1999.

Volker Halbach. *Axiomatic Theories of Truth*. Cambridge University Press, 2011.

Volker Halbach and Leon Horsten. Computational structuralism. *Philosophia Mathematica*, 13(2):174–186, 2005.

Volker Halbach and Leon Horsten. Axiomatizing kripke's theory of truth. *Journal of Symbolic Logic*, 71(2):677–712, 2006.

Joel Hamkins and Ruizhi Yang. Satisfaction is not absolute. *ArXiv e-prints*, 2013. arXiv:1312.0670.

Geoffrey Hellman. Against 'absolutely everything'! In Agustín Rayo and Gabriel Uzquiano, editors, *Absolute Generality*, pages 75–97. Oxford University Press, 2006.

Eric Hiddleston. Second-order properties and three varieties of functionalism. *Philosophical Studies*, 153(3):397–415, 2011.

Leon Horsten. The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In P. Cartois, editor, *The Many Problems of Realism*, pages 173–187. Tilburg University Press, 1995.

Leon Horsten. *The Tarskian Turn. Deflationism and Axiomatic Truth*. MIT Press, 2011.

Leon Horsten and Volker Halbach. Norms for theories of reflexive truth. In Kentaro Fujimoto, José Martínez Fernández, Henri Galinon, and Theodora Achourioti, editors, *Unifying the Philosophy of Truth*. Springer Verlag, 2015.

Leon Horsten and Graham Leigh. Truth is simple. *Mind*, 126(501):195–232, 2017.

Paul Horwich. *Truth*. Clarendon Press, 1998.

Daniel Isaacson. Arithmetical truth and hidden higher-order concepts. In The Paris Logic Group, editor, *Logic Colloquium '85*, volume 122 of *Studies in Logic and the Foundations of Mathematics*, pages 147–169. Elsevier, 1987.

Mark Johnston. How to speak of the colors. *Philosophical Studies*, 68(3):221–263, 1992.

Vladimir Kanovei and Michael Reeken. A nonstandard proof of the jordan curve theorem. *Real Analysis Exchange*, 24(1):161–170, 1998.

Richard Kaye. *Models of Peano arithmetic*, volume 15 of *Oxford Logic Guides*. Oxford University Press, 1991.

Richard Kaye. *The mathematics of logic*. Cambridge University Press, Cambridge, 2007. A guide to completeness theorems and their applications.

Richard Kaye. *Models of Peano arithmetic*. Second Edition (unpublished), please contact the author for a copy, 2012.

Jeffrey Ketland. Deflationism and tarski's paradise. *Mind*, 108(429):69–94, 1999.

Stephen Kleene. *Introduction to Metamathematics*. North Holland, 1952.

Henryk Kotlarski. Bounded induction and satisfaction classes. *Mathematical Logic Quarterly*, 32:531–544, 1986.

Henryk Kotlarski. Full satisfaction classes: a survey. *Notre Dame Journal of Formal Logic*, 32(4):573–579, 1991.

Henryk Kotlarski, Stanisław Krajewski, and Alistair Lachlan. Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin*, 24(3):283–293, 1981.

Saul Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72(19):690–716, 1975.

Wolfgang Künne. *Conceptions of Truth*. Oxford University Press, 2003.

George Leibman. A nonstandard proof of the fundamental theorem of algebra. *The American Mathematical Monthly*, 112(8):705–712, 2005.

Hannes Leitgeb. What theories of truth should be like (but cannot be). *Philosophy Compass*, 2(2):276–290, 2007.

Mateusz Łełyk. *Axiomatic Theories of Truth, Bounded Induction and Reflection Principles*. PhD thesis, Uniwersytet Warszawski, 2017.

David Lewis. *On the Plurality of Worlds*. Wiley-Blackwell, 1986.

David Liggins. Deflationism, conceptual explanation, and the truth asymmetry. *Philosophical Quarterly*, 66(262):84–101, 2016.

Martin Löb. Solution of a problem of Leon Henkin. *The Journal of Symbolic Logic*, 20(2):115–118, 1955.

Michael Lynch. Précis to *True to Life*, and replies to commentators. *Philosophical Books*, 46:289–291, 331–342, 2005.

Michael Lynch. *Truth as One and Many*. Clarendon Press, 2009.

Michael Lynch. Three questions for truth pluralism. In Nikolaj Pedersen and Cory Wright, editors, *Truth and Pluralism: Current Debates*, pages 21–41. Oxford University Press, 2013.

Paolo Mancosu. Explanation in mathematics. In Edward Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2015.

Vann McGee. Maximal consistent sets of instances of tarski's schema (T). *Journal of Philosophical Logic*, 21(3):235–241, 1992.

Colin McGinn. *Logical Properties: Identity, Existence, Predication, Necessity, Truth*. Oxford University Press, 2000.

Cheryl Misak. The pragmatist theory of truth. In Michael Glanzberg, editor, *The Oxford Handbook of Truth*. Oxford University Press, 2018.

George Moore. The nature of judgment. *Mind*, 8(2):176–193, 1899.

Carlo Nicolai. Deflationary truth and the ontology of expressions. *Synthese*, 192 (12):4031–4055, 2015.

Charles Parsons. The liar paradox. *Journal of Philosophical Logic*, 3(4):381–412, 1974.

Douglas Patterson. What is a correspondence theory of truth? *Synthese*, 137(3): 421–444, 2003.

Nikolaj Pedersen. Stabilizing alethic pluralism. *Philosophical Quarterly*, 60(238): 92–108, 2010.

Nikolaj Pedersen. Pluralism ×3: Truth, logic, metaphysics. *Erkenntnis*, 79(S2): 259–277, 2014.

Nikolaj Pedersen and Cory Wright. Pluralism about truth as alethic disjunctivism. In Nikolaj Pedersen and Cory Wright, editors, *Truth and Pluralism: Current Debates*, pages 82–112. Oxford University Press, 2013.

Lavinia Picollo. Reference in arithmetic. *Review of Symbolic Logic*, 11(3):573–603, 2018.

Lavinia Picollo and Thomas Schindler. Disquotation and infinite conjunctions. *Erkenntnis*, 83(5):899–928, 2018.

Dag Prawitz. Meaning approached via proofs. *Synthese*, 148(3):507–524, 2006.

Hilary Putnam. *Meaning and the Moral Sciences*. Routledge and Kegan Paul, 1978.

Willard Quine. *Word and Object*. MIT Press, 1960.

Willard Quine. *Philosophy of Logic*. Harvard University Press, 1986.

Panu Raatikainen. Problems of deflationism. In Tuomo Aho and Ahti-Veikko Pietarinen, editors, *Truth and Games in Logic and Language*, pages 175–185. 2006.

Frank Ramsey. Facts and propositions. *Proceedings of the Aristotelian Society*, 7 (1):153–170, 1927.

Agustín Rayo and Gabriel Uzquiano. Introduction. In Agustín Rayo and Gabriel Uzquiano, editors, *Absolute Generality*, pages 1–19. Oxford University Press, 2006.

Abraham Robinson. On languages which are based on non-standard arithmetic. *Nagoya Mathematical Journal*, 22:83–117, 1963.

Abraham Robinson and Peter Roquette. On the finiteness theorem of siegel and mahler concerning diophantine equations. *Journal of Number Theory*, 7(2):121–176, 1975.

Bertrand Russell. Meinong's theory of complexes and assumptions (III.). *Mind*, 13(52):509–524, 1904.

Bertrand Russell. Mathematical logic as based on the theory of types. *American Journal of Mathematics*, 30(3):222–262, 1908.

Mark Sainsbury. Review: Crispin Wright: Truth and objectivity. *Philosophy and Phenomenological Research*, 56(4):899–904, 1996.

Stephen Schiffer. *The Things We Mean*. Oxford University Press, 2003.

Stewart Shapiro. Proof and truth: Through thick and thin. *Journal of Philosophy*, 95(10):493–521, 1998.

Stewart Shapiro. Truth, function and paradox. *Analysis*, 71(1):38–44, 2011.

Gila Sher. Functional pluralism. *Philosophical Books*, 46(4):311–330, 2005.

Gila Sher. Tarski's thesis. In Douglas Patterson, editor, *New Essays on Tarski and Philosophy*, pages 300–339. Oxford University Press, 2008.

Alan Sidelle. *Necessity, Essence, and Individuation: A Defense of Conventionalism*. Cornell University Press, 1989.

Keith Simmons. *Semantic Singularities: Paradoxes of Reference, Predication, and Truth*. Oxford University Press, 2018.

Peter Smith. *An Introduction to Gödel's Theorems*. Cambridge Introductions to Philosophy. Cambridge University Press, 2007.

Scott Soames. *Understanding Truth*. Oxford University Press, 1998.

Johannes Stern. Modality and axiomatic theories of truth I: Friedman-Sheard. *The Review of Symbolic Logic*, 7(2):273–298, 2014a.

Johannes Stern. Modality and axiomatic theories of truth II: Kripke-Feferman. *The Review of Symbolic Logic*, 7(2):299–318, 2014b.

Peter Strawson. Truth. *Analysis*, 9(6):83–97, 1948.

Andrea Strollo. Deflationism and the invisible power of truth. *Dialectica*, 67(4): 521–543, 2013.

Andrea Strollo. How simple is the simplicity of truth? Reconciling the mathematics and the metaphysics of truth. In Fabio Bacchini, Stefano Caputo, and Massimo Dell'Utri, editors, *New Frontiers in Truth*, pages 161–175. Cambridge Scholars Publishing, 2014.

Christine Tappolet. Truth pluralism and many-valued logics: A reply to Beall. *Philosophical Quarterly*, 50(200):382–385, 2000.

Alfred Tarski. *Logic, semantics, metamathematics. Papers from 1923 to 1938*. Oxford at the Clarendon Press, 1956. Translated by J. H. Woodger.

Alfred Tarski. What are logical notions? *History and Philosophy of Logic*, 7(2): 143–154, 1986. Edited by John Corcoran.

Neil Tennant. Deflationism and the Gödel phenomena. *Mind*, 111(443):551–582, 2002.

Paul Thagard. Coherence, truth, and the development of scientific knowledge. *Philosophy of Science*, 74(1):28–47, 2007.

Amie Thomasson. *Ontology Made Easy*. Oxford University Press, 2015.

Peter van Inwagen. *Material Beings*. Cornell University Press, 1990.

Albert Visser. Semantics and the liar paradox. *Handbook of Philosophical Logic*, 4(1):617–706, 1989.

Ralph Walker. The coherence theory. In Michael Lynch, editor, *The Nature of Truth: Classic and Contemporary Perspectives*, pages 123–158. MIT Press, 2001.

Bartosz Wcisło and Mateusz Łełyk. Notes on bounded induction for the compositional truth predicate. *The Review of Symbolic Logic*, 10(3):455–480, 2017.

Timothy Williamson. Everything. *Philosophical Perspectives*, 17(1):415–465, 2003.

Cory Wright. On the functionalization of pluralist approaches to truth. *Synthese*, 145(1):1–28, 2005.

Crispin Wright. Truth: A traditional debate reviewed. *Canadian Journal of Philosophy*, 28(S1):31–74, 1998.

Crispin Wright. Minimalism, deflationism, pragmatism, pluralism. In Michael Lynch, editor, *The Nature of Truth: Classic and Contemporary Perspectives*, pages 751–787. MIT Press, 2001.

Crispin Wright. A plurality of pluralisms. In Nikolaj Pedersen and Cory Wright, editors, *Truth and Pluralism: Current Debates*, pages 123–153. Oxford University Press, 2013.

Jeremy Wyatt. Domains, plural truth, and mixed atomic propositions. *Philosophical Studies*, 166(S1):225–236, 2013.

Jeremy Wyatt. The many (yet few) faces of deflationism. *The Philosophical Quarterly*, 66(263):362–382, 2016.

Jeremy Wyatt and Michael Lynch. From one to many: Recent work on truth. *American Philosophical Quarterly*, 53(4):323–340, 2016.

Stephen Yablo. Truth and reflection. *Journal of Philosophical Logic*, 14(3):297–349, 1985.

Edward Zalta. *Abstract Objects: An Introduction to Axiomatic Metaphysics*. D. Reidel Publishing Company, 1983.